Gene Profiling of Human Embryonic Stem Cells

Related Application

This application claims priority to United States Provisional Application 60/542,451, filed February 6, 2004, the disclosure of which is hereby incorporated by reference in its entirety.

Background of the Invention

Pluripotential stem cells isolated from either adult or embryonic sources have generated intense scientific and public interest over the last several years. Such cells likely hold the key to an array of novel therapies for diseases and injuries which affect virtually any tissue. Given the financial and personal toll exacted by diseases such as diabetes, stroke, Alzheimer's disease, Parkinson's disease, cardiovascular disease, arthritis, and the like, the need for further research aimed at increasing our understanding of stem cell biology is clear.

One of the limitations of the current state of the stem cell art is the relative scarcity of our understanding of the molecular nature of stem cells. Furthermore, since the focus of many stem cell studies has been the overt behavior of the cells (e.g., what do the cells look like under various conditions), our understanding of the molecular underpinnings driving stem cell development, proliferation, differentiation, and survival is incomplete. Perhaps the most serious impact of our limited molecular understanding of stem cells is the difficulty of identifying stem cells and following their movement and fate over time. A more detailed molecular understanding of stem cells would provide specific markers of various stages of stem cell development. Such markers could be used, for example, to (i) identify stem cells from amongst a heterogeneous population of cells in culture, (ii) identify stem cells from amongst a heterogeneous population of cells in vivo, (iii) purify stem cells from amongst a heterogeneous population of cells, (iv) establish assays based on the identification of stem cells, (v) establish assays based on identification of agents which induce a stem cell like fate, and (vi) monitor the process or progress of cellular differentiation.

These gaps in our knowledge of stem cells, particularly of the molecular nature of stem cells, have consequences that extend beyond mere scientific

frustration. Although these points are equally applicable to embryonic and adult stem cells, we have focused our attention on embryonic stem cells. We note however, that progress in the adult stem cell field has been equally hampered by our limited molecular understanding of and ability to unequivocally identify stem cells from amongst heterogeneous populations of cells. In fact, given that the term "adult stem cell" encompasses many diverse cell types which differ in morphology, gene expression, growth characteristics, differentiation potential, place of origin, etc., the issues which we herein address with respect to embryonic stem cells must be faced in analyzing each of the various adult stem cell populations.

Summary of the Invention

Embryonic stem cells are an area of intense research, and there exists a substantial need for improved methods of identifying embryonic stem cells. Such improved methods of identification have applications in basic research, as well as in the design of improved drug screening assays, improved therapeutic regimens, and improved isolation and purification methods. The present invention aims to satisfy the need in the art for improved methods and compositions for identifying embryonic stem cells.

In a first aspect, the present invention provides four novel markers of embryonic stem cells. These novel markers can be used in the design of probes and primers to identify embryonic stem cells.

In one embodiment, the probe/primer comprises a nucleic acid that hybridizes under stringent conditions, including a wash step of 0.2X SSC at 65 °C, to a nucleic acid represented in SEQ ID NO: 3 or a complement thereof. In another embodiment, the probe/primer comprises a nucleic acid that hybridizes under stringent conditions, including a wash step of 0.2X SSC at 65 °C, to a nucleic acid represented in SEQ ID NO: 5 or a complement thereof. In another embodiment, the probe/primer comprises a nucleic acid that hybridizes under stringent conditions, including a wash step of 0.2X SSC at 65 °C, to a nucleic acid represented in SEQ ID NO: 7 or a complement thereof. In yet another embodiment, the probe/primer comprises a nucleic acid that hybridizes under stringent conditions, including a wash

step of 0.2X SSC at 65 °C, to a nucleic acid represented in SEQ ID NO: 9 or a complement thereof.

In one embodiment, the embryonic stem cells are mammalian embryonic stem cells. In another embodiment, the mammalian embryonic stem cells are human embryonic stem cells.

In one embodiment, the probe/primer is detectably labeled.

In yet another embodiment, the probe/primer comprises all or a portion of a nucleic acid represented in any of SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7, or SEQ ID NO: 9. In preferred embodiments, the portion includes at least 10, 12, 15, 18, 20, 22, 25, 50, 75, or greater than 75 nucleotides.

In yet another embodiment, the invention contemplates a composition comprising more than one embryonic stem cell specific probe/primer. The invention contemplates compositions comprising two, three, four, or more than four of the foregoing probe/primers.

In a second aspect, the present invention provides the use of a probe/primer in the manufacture of a composition for identifying an embryonic stem cell. In one embodiment, the probe/primer comprises a nucleic acid that hybridizes under stringent conditions, including a wash step of 0.2X SSC at 65 °C, to a nucleic acid represented in SEQ ID NO: 3 or a complement thereof. In another embodiment, the probe/primer comprises a nucleic acid that hybridizes under stringent conditions, including a wash step of 0.2X SSC at 65 °C, to a nucleic acid represented in SEQ ID NO: 5 or a complement thereof. In another embodiment, the probe/primer comprises a nucleic acid that hybridizes under stringent conditions, including a wash step of 0.2X SSC at 65 °C, to a nucleic acid represented in SEQ ID NO: 7 or a complement thereof. In yet another embodiment, the probe/primer comprises a nucleic acid that hybridizes under stringent conditions, including a wash step of 0.2X SSC at 65 °C, to a nucleic acid represented in SEQ ID NO: 9 or a complement thereof.

In one embodiment, the embryonic stem cells are mammalian embryonic stem cells. In another embodiment, the mammalian embryonic stem cells are human embryonic stem cells.

In one embodiment, the probe/primer is detectably labeled.

In yet another embodiment, the probe/primer comprises all or a portion of a nucleic acid represented in any of SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7, or SEQ ID NO: 9. In preferred embodiments, the portion includes at least 10, 12, 15, 18, 20, 22, 25, 50, 75, or greater than 75 nucleotides.

In yet another embodiment, the invention contemplates a composition comprising more than one embryonic stem cell specific probe/primer. The invention contemplates compositions comprising two, three, four, or more than four of the foregoing probe/primers.

In a third aspect, the present invention provides a method of determining whether a cell is an embryonic stem cell. The method comprises the following steps: contacting a population of cells with an embryonic stem cell specific probe/primer and identifying the one or more cells containing a nucleic acid sequence to which said probe/primer hybridizes. The one or more cells that contain a nucleic acid sequence to which said probe/primer hybridizes is determined to be an embryonic stem cell. In one embodiment, the probe/primer comprises a nucleic acid that hybridizes under stringent conditions, including a wash step of 0.2X SSC at 65 °C, to a nucleic acid represented in SEQ ID NO: 3 or a complement thereof. In another embodiment, the probe/primer comprises a nucleic acid that hybridizes under stringent conditions, including a wash step of 0.2X SSC at 65 °C, to a nucleic acid represented in SEQ ID NO: 5 or a complement thereof. In another embodiment, the probe/primer comprises a nucleic acid that hybridizes under stringent conditions, including a wash step of 0.2X SSC at 65 °C, to a nucleic acid represented in SEQ ID NO: 7 or a complement thereof. In yet another embodiment, the probe/primer comprises a nucleic acid that hybridizes under stringent conditions, including a wash step of 0.2X SSC at 65 °C, to a nucleic acid represented in SEQ ID NO: 9 or a complement thereof.

In one embodiment, the embryonic stem cells are mammalian embryonic stem cells. In another embodiment, the mammalian embryonic stem cells are human embryonic stem cells.

In one embodiment, the probe/primer is detectably labeled.

In yet another embodiment, the probe/primer comprises all or a portion of a nucleic acid represented in any of SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7, or

SEQ ID NO: 9. In preferred embodiments, the portion includes at least 10, 12, 15, 18, 20, 22, 25, 50, 75, or greater than 75 nucleotides.

In yet another embodiment, the method comprising determining whether a cell is an embryonic stem cell using more than one of the foregoing probe/primers. The invention contemplates a method comprising two, three, four, or more than four probe/primers.

In a fourth aspect, the present invention provides the use of primer pairs in the manufacture of a composition for identifying embryonic stem cells. In one embodiment, the primer pair comprises SEQ ID NO: 13 and SEQ ID NO: 14. In another embodiment, the primer pair comprises SEQ ID NO: 15 and SEQ ID NO: 16. In another embodiment, the primer pair comprises SEQ ID NO: 19 and SEQ ID NO: 20. In yet another embodiment, the primer pair comprises SEQ ID NO: 11 and SEQ ID NO: 12.

In one embodiment, the embryonic stem cells are mammalian embryonic stem cells. In another embodiment, the mammalian embryonic stem cells are human embryonic stem cells.

In a fifth aspect, the present invention provides the use of a primer pair in the manufacture of a composition for amplifying an embryonic stem cells marker. In one embodiment, the primer pair comprises SEQ ID NO: 13 and SEQ ID NO: 14. In another embodiment, the primer pair comprises SEQ ID NO: 15 and SEQ ID NO: 16. In another embodiment, the primer pair comprises SEQ ID NO: 19 and SEQ ID NO: 20. In yet another embodiment, the primer pair comprises SEQ ID NO: 11 and SEQ ID NO: 12.

In one embodiment, the embryonic stem cells are mammalian embryonic stem cells. In another embodiment, the mammalian embryonic stem cells are human embryonic stem cells.

In a sixth aspect, the present invention provides nucleic acids comprising embryonic stem cell markers operably linked to transcriptional regulatory sequences so as to render the nucleic acid suitable for use as an expression vector. The present invention further provides expression vectors comprising these embryonic stem cell markers, host cells expressing polypeptides encoded by these markers, and methods of producing recombinant polypeptides.

In a seventh aspect, the present invention provides a method of promoting an embryonic stem cell phenotype in a cell. In one embodiment, the method comprises expressing one embryonic stem cell specific gene or protein in a cell. In another embodiment, the method comprises expressing more than one (e.g., two, three, or four) embryonic stem cell specific gene or protein in a cell.

In an eighth aspect, the present invention provides a method of promoting an embryonic stem cell phenotype in a cell comprising administering an amount of an agent effective to increase the expression of an embryonic stem cell marker. In one embodiment, the embryonic stem cell marker comprises a nucleic acid sequence represented in SEQ ID NO: 3. In another embodiment, the embryonic stem cell marker comprises a nucleic acid sequence represented in SEQ ID NO: 5. In another embodiment, the embryonic stem cell marker comprises a nucleic acid sequence represented in SEQ ID NO: 7. In still another embodiment, the embryonic stem cells marker comprises a nucleic acid sequence represented in SEQ ID NO: 9.

In one embodiment, the method comprises administering more than one agent. In another embodiment, the method comprises administering one or more agents that increase the expression of more than one embryonic stem cell markers.

In a ninth aspect, the present invention provides a method of inhibiting an embryonic stem cell phenotype in a cell comprising administering an amount of an agent effective to decrease the expression of an embryonic stem cell marker. In one embodiment, the embryonic stem cells marker comprises a nucleic acid sequence represented in SEQ ID NO: 3. In another embodiment, the embryonic stem cell marker comprises a nucleic acid sequence represented in SEQ ID NO: 5. In another embodiment, the embryonic stem cell marker comprises a nucleic acid sequence represented in SEQ ID NO: 7. In still another embodiment, the embryonic stem cells marker comprises a nucleic acid sequence represented in SEQ ID NO: 9.

In one embodiment, the method comprises administering more than one agent. In another embodiment, the method comprises administering one or more agents that decrease the expression of more than one embryonic stem cell markers.

In a tenth aspect, the present invention provides a method of purifying embryonic stem cells based on the expression of one or more (e.g., one, two, three, or four) embryonic stem cell markers or based on the expression of a reporter gene

regulated by a regulatory region of an embryonic stem cell marker. The embryonic stem cell marker corresponds to all or a portion of the sequence represented in SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7, or SEQ ID NO: 9, or to a regulatory region which endogenously regulates the expression of any of the foregoing.

In one embodiment, the method comprises detecting the expression of a reporter construct. The reporter construct comprises all or a portion of the promoter of an embryonic stem cell specific marker represented in SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7, or SEQ ID NO: 9. The reporter can be detected in living cells and expression of this reporter allows the identification and purification of cells which express an embryonic stem cell specific marker (i.e., cells which express a reporter gene operably linked to a portion of a promoter of an embryonic stem cell specific gene sufficient to regulate expression of the stem cell specific gene in cells.

In another embodiment, the method of detecting the embryonic stem cell marker comprises a probe/primer comprising a nucleic acid that hybridizes under stringent conditions, including a wash step of 0.2X SSC at 65 °C, to a nucleic acid represented in SEQ ID NO: 3 or a complement thereof. In another embodiment, the embryonic stem cell marker comprises a probe/primer comprising a nucleic acid that hybridizes under stringent conditions, including a wash step of 0.2X SSC at 65 °C, to a nucleic acid represented in SEQ ID NO: 5 or a complement thereof. In another embodiment, the embryonic stem cell marker comprises a probe/primer comprising a nucleic acid that hybridizes under stringent conditions, including a wash step of 0.2X SSC at 65 °C, to a nucleic acid represented in SEQ ID NO: 7 or a complement thereof. In yet another embodiment, the embryonic stem cell marker comprises a probe/primer comprising a nucleic acid that hybridizes under stringent conditions, including a wash step of 0.2X SSC at 65 °C, to a nucleic acid represented in SEQ ID NO: 9 or a complement thereof.

In an eleventh aspect, the present invention provides a method of enriching a population of cells to increase the proportion of embryonic stem cells using an embryonic stem cell marker or based on the expression of a reporter gene regulated by a regulatory region of an embryonic stem cell marker. The embryonic stem cell marker corresponds to all or a portion of the sequence represented in SEQ ID NO: 3,

SEQ ID NO: 5, SEQ ID NO: 7, or SEQ ID NO: 9, or to a regulatory region which endogenously regulates the expression of any of the foregoing.

In one embodiment, the method comprises detecting the expression of a reporter construct. The reporter construct comprises all or a portion of the promoter of a stem cell specific marker represented in SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7, or SEQ ID NO: 9. The reporter can be detected in living cells and expression of this reporter allows the identification and purification of cells which express a stem cell specific marker (i.e., cells which express a reporter gene operably linked to a portion of a promoter of a stem cell specific gene sufficient to regulate expression of the stem cell specific gene in cells.

In another embodiment, the method of enriching comprises a probe/primer comprising a nucleic acid that hybridizes under stringent conditions, including a wash step of 0.2X SSC at 65 °C, to a nucleic acid represented in SEQ ID NO: 3 or a complement thereof. In another embodiment, the embryonic stem cell marker comprises a probe/primer comprising a nucleic acid that hybridizes under stringent conditions, including a wash step of 0.2X SSC at 65 °C, to a nucleic acid represented in SEQ ID NO: 5 or a complement thereof. In another embodiment, the embryonic stem cell marker comprises a probe/primer comprising a nucleic acid that hybridizes under stringent conditions, including a wash step of 0.2X SSC at 65 °C, to a nucleic acid represented in SEQ ID NO: 7 or a complement thereof. In yet another embodiment, the embryonic stem cell marker comprises a probe/primer comprising a nucleic acid that hybridizes under stringent conditions, including a wash step of 0.2X SSC at 65 °C, to a nucleic acid represented in SEQ ID NO: 9 or a complement thereof.

In a twelfth aspect, the present invention provides cells engineered to recombinantly express a reporter construct including all or a portion of a regulatory region (i.e., promoter or enhancer sequences) of a stem cell specific marker represented in SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7, or SEQ ID NO: 9. The reporter construct comprises all or a portion of said regulatory region and a reporter gene which encodes a detectable marker.

In a thirteenth aspect, the present invention provides antibodies that are specifically immunoreactive with a protein comprising an amino acid sequence

encodable by a nucleic acid sequence represented in SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7, or SEQ ID NO: 9.

In one embodiment, the antibody is a monoclonal antibody. In another embodiment, the antibody is a polyclonal antibody.

In one embodiment, the antibody is used in a method of identifying, characterizing, and/or purifying a cell with an embryonic stem cell phenotype.

In a fourteenth aspect, the present invention provides a method of identifying embryonic stem cells, comprising examining the expression of one or more of the markers presented in Table 1.

In one embodiment, the method comprises examining 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 66, or 67 of the markers presented in Table 1.

The practice of the present invention will employ, unless otherwise indicated, conventional techniques of cell biology, cell culture, molecular biology, transgenic biology, microbiology, virology, recombinant DNA, and immunology, which are within the skill of the art. Such techniques are described in the literature. See, for example, Molecular Cloning: A Laboratory Manual, 3rd Ed., ed. by Sambrook and Russell (Cold Spring Harbor Laboratory Press: 2001); the treatise, Methods In Enzymology (Academic Press, Inc., N.Y.); Using Antibodies, Second Edition by Harlow and Lane, Cold Spring Harbor Press, New York, 1999; Current Protocols in Cell Biology, ed. by Bonifacino, Dasso, Lippincott-Schwartz, Harford, and Yamada, John Wiley and Sons, Inc., New York, 1999.

Other features and advantages of the invention will be apparent from the following detailed description, and from the claims.

Brief Description of the Drawings

Figure 1 shows a graphical representation of gene profile analysis of human embryonic stem cell lines, in comparison to either other embryonic stem cell lines, embryoid bodies, or other somatic cell types. Figure 1A shows a representation which demonstrates the degree of similarity in gene expression between human ES cells and various other tissues. Results are displayed as a branched tree representing the degree of similarity as a function of branch height. The degree of similarity was

determined by hierarchical clustering of normalized expression values obtained in ES cells, EBs, and over 100 tissue samples. Figure 1B shows a representation demonstrating the degree of similarity between two independently derived ES cell lines (cell lines X and Y) and EBs (derived from one of these cell lines) following a number of days of differentiation (2 days, 10 days, and 30 days). Hierarchical clustering is as in Figure 1A.

Figure 2 shows the expression of identified ES specific genes. The ES specific expression of the five identified ES specific genes was confirmed by RT-PCR. Briefly, the expression of the five ES specific genes in either ES cells or in day 20 embryoid bodies was verified by RT-PCR. Amplification of GPDH served as a control. The RT-PCR analysis confirms that OCT4 and the four novel ES specific markers are specifically expressed in ES cells.

Figure 3 provides a summary of the analysis performed on the ES specific nucleic acid represented in SEQ ID NO: 3.

Figure 4 provides a summary of the analysis performed on the ES specific nucleic acid represented in SEQ ID NO: 5.

Figure 5 provides a summary of the analysis performed on the ES specific nucleic acid represented in SEQ ID NO: 7.

Figure 6 provides a summary of the analysis performed on the ES specific nucleic acid represented in SEQ ID NO: 9.

Figure 7 shows that differentiation within embryoid bodies recapitulates a normal pattern of gene expression observed during development. The expression patterns of genes involved with Nodal signaling were analyzed in ES cells, and EBs of various ages. Figure 7A provides a graphical representation of the expression patterns of nodal, leftyA, leftyB, and pitx2, as determined by micro-array analysis. This pattern recapitulates the temporal pattern observed during development. Figure 7B provides RT-PCR analysis that validated the results obtained from the micro-array analysis. We note that GPDH served as a control in the RT-PCR experiments.

Table 1 provides a list of markers that are expressed preferentially in embryonic stem cells in comparison to embryoid bodies.

Detailed Description of the Invention

(i) Overview

Although the past several years has seen tremendous advances in the field of stem cell research, additional advances require improved methods of identifying, isolating, and purifying embryonic stem cells. Such improved methods and compositions will facilitate the types of additional studies that will increase our understanding of the basic biology underlying embryonic stem cells.

Currently, the paucity of embryonic stem cell markers limits improved methods of identifying and purifying stem cells, as well as limits methods of enriching mixed population of embryonic stem cells. The limitations serve to slow the important developments in stem cell research needed to move this work from the laboratory to the clinic. The present invention offers significant improvements over the prior art by providing novel markers of embryonic stem cells which are expressed in embryonic stem cells and germ cells to a substantially greater degree than in even other related cell populations such as embryoid bodies. Based on the discovery of four novel markers, the present invention provides compositions designed to detect expression of these novel markers (i.e., probes/primers), and methods of using these novel markers to identify, characterize, and purify embryonic stem cells. Accordingly, the present invention provides methods and compositions aimed at facilitating the study and use of embryonic stem cells, as well as methods and compositions which improve our molecular understanding of the biology underlying the embryonic stem cell phenotype.

Additionally, the present invention provides an extensive list of markers that are expressed preferentially in embryonic stem cells in comparison to embryoid bodies. Analysis of one or more of these markers allows for improved methods of identifying and characterizing embryonic stem cells in comparison to more differentiated cell types.

(ii) Definitions

For convenience, certain terms employed in the specification, examples, and appended claims are collected here. Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs.

The articles "a" and "an" are used herein to refer to one or to more than one (i.e., to at least one) of the grammatical object of the article. By way of example, "an element" means one element or more than one element.

As used herein, "protein" is a polymer consisting essentially of any of the 20 amino acids. Although "polypeptide" is often used in reference to relatively large polypeptides, and "peptide" is often used in reference to small polypeptides, usage of these terms in the art overlaps and is varied.

The terms "peptide(s)", "protein(s)" and "polypeptide(s)" are used interchangeably herein.

The terms "polynucleotide sequence" and "nucleotide sequence" are also used interchangeably herein.

"Recombinant," as used herein, means that a protein is derived from a prokaryotic or eukaryotic expression system.

The term "wild type" refers to the naturally-occurring polynucleotide sequence encoding a protein, or a portion thereof, or protein sequence, or portion thereof, respectively, as it normally exists *in vivo*.

The term "mutant" refers to any change in the genetic material of an organism, in particular a change (i.e., deletion, substitution, addition, or alteration) in a wildtype polynucleotide sequence or any change in a wildtype protein sequence. The term "variant" is used interchangeably with "mutant". Although it is often assumed that a change in the genetic material results in a change of the function of the protein, the terms "mutant" and "variant" refer to a change in the sequence of a wildtype protein regardless of whether that change alters the function of the protein (e.g., increases, decreases, imparts a new function), or whether that change has no effect on the function of the protein (e.g., the mutation or variation is silent).

As used herein, the term "nucleic acid" refers to polynucleotides such as deoxyribonucleic acid (DNA), and, where appropriate, ribonucleic acid (RNA). The term should also be understood to include, as equivalents, analogs of either RNA or DNA made from nucleotide analogs, and, as applicable to the embodiment being described, single (sense or antisense) and double-stranded polynucleotides.

As used herein, the term "gene" or "recombinant gene" refers to a nucleic acid comprising an open reading frame encoding a polypeptide, including both exon and (optionally) intron sequences.

As used herein, the term "vector" refers to a nucleic acid molecule capable of transporting another nucleic acid to which it has been linked. Preferred vectors are those capable of autonomous replication and/or expression of nucleic acids to which they are linked. Vectors capable of directing the expression of genes to which they are operatively linked are referred to herein as "expression vectors".

A polynucleotide sequence (DNA, RNA) is "operatively linked" to an expression control sequence when the expression control sequence controls and regulates the transcription and translation of that polynucleotide sequence. The term "operatively linked" includes having an appropriate start signal (e.g., ATG) in front of the polynucleotide sequence to be expressed, and maintaining the correct reading frame to permit expression of the polynucleotide sequence under the control of the expression control sequence, and production of the desired polypeptide encoded by the polynucleotide sequence.

"Transcriptional regulatory sequence" is a generic term used throughout the specification to refer to nucleic acid sequences, such as initiation signals, enhancers, and promoters, which induce or control transcription of protein coding sequences with which they are operably linked. In some examples, transcription of a recombinant gene is under the control of a promoter sequence (or other transcriptional regulatory sequence) which controls the expression of the recombinant gene in a cell-type in which expression is intended. It will also be understood that the recombinant gene can be under the control of transcriptional regulatory sequences which are the same or which are different from those sequences which control transcription of the naturally-occurring form of a protein.

As used herein, the term "tissue-specific promoter" means a nucleic acid sequence that serves as a promoter, i.e., regulates expression of a selected nucleic acid sequence operably linked to the promoter, and which affects expression of the selected nucleic acid sequence in specific cells of a tissue, such as cells of neural origin, e.g. neuronal cells. The term also covers so-called "leaky" promoters, which

regulate expression of a selected nucleic acid primarily in one tissue, but cause expression in other tissues as well.

"Homology" and "identity" are used synonymously throughout and refer to sequence similarity between two peptides or between two nucleic acid molecules. Homology can be determined by comparing a position in each sequence which may be aligned for purposes of comparison. When a position in the compared sequence is occupied by the same base or amino acid, then the molecules are homologous or identical at that position. A degree of homology or identity between sequences is a function of the number of matching or homologous positions shared by the sequences.

A "chimeric protein" or "fusion protein" is a fusion of a first amino acid sequence encoding a polypeptide with a second amino acid sequence defining a domain (e.g. polypeptide portion) foreign to and not substantially homologous with any domain of the first polypeptide. A chimeric protein may present a foreign domain which is found (albeit in a different protein) in an organism which also expresses the first protein, or it may be an "interspecies", "intergenic", etc. fusion of protein structures expressed by different kinds of organisms.

The "non-human animals" of the invention include mammals such as rats, mice, rabbits, sheep, cats, dogs, cows, pigs, and non-human primates.

The term "isolated" as used herein with respect to nucleic acids, such as DNA or RNA, refers to molecules separated from other DNAs, or RNAs, respectively, that are present in the natural source of the macromolecule. For example, an isolated nucleic acid preferably includes no more than 10 kilobases (kb) of nucleic acid sequence which naturally immediately flanks the gene in genomic DNA, more preferably no more than 5kb of such naturally occurring flanking sequences, and most preferably less than 1.5kb of such naturally occurring flanking sequence. The term isolated as used herein also refers to a nucleic acid or peptide that is substantially free of cellular material, or culture medium when produced by recombinant DNA techniques, or chemical precursors or other chemicals when chemically synthesized. Moreover, an "isolated nucleic acid" is meant to include nucleic acid fragments which are not naturally occurring as fragments and would not be found in the natural state.

As used herein, "proliferating" and "proliferation" refer to cells undergoing mitosis.

As used herein the term "animal" refers to mammals, preferably mammals such as humans. Likewise, a "patient" or "subject" to be treated by the method of the invention can mean either a human or non-human animal.

"Differentiation" in the present context means the formation of cells expressing markers known to be associated with cells that are more specialized and closer to becoming terminally differentiated cells incapable of further division or differentiation.

The term "progenitor cell" is used synonymously with "stem cell". Both terms refer to an undifferentiated cell which is capable of proliferation and giving rise to more progenitor cells having the ability to generate a large number of mother cells that can in turn give rise to differentiated, or differentiable daughter cells. In a preferred embodiment, the term progenitor or stem cell refers to a generalized mother cell whose descendants (progeny) specialize, often in different directions, by differentiation, e.g., by acquiring completely individual characters, as occurs in progressive diversification of embryonic cells and tissues. Cellular differentiation is a complex process typically occurring through many cell divisions. A differentiated cell may derive from a multipotent cell which itself is derived from a multipotent cell, and so on. While each of these multipotent cells may be considered stem cells, the range of cell types each can give rise to may vary considerably. Some differentiated cells also have the capacity to give rise to cells of greater developmental potential. Such capacity may be natural or may be induced artificially upon treatment with various factors.

The term "embryonic stem cell" is used to refer to the pluripotent stem cells of the inner cell mass of the embryonic blastocyst (see US Patent Nos. 5843780, 6200806). Such cells can similarly be obtained from the inner cell mass of blastocysts derived from somatic cell nuclear transfer (see, for example, US Patent Nos 5945577, 5994619, 6235970). The distinguishing characteristics of an embryonic stem cell define an embryonic stem cell phenotype. Accordingly, a cell has the phenotype of an embryonic stem cell if it possesses one or more of the unique characteristics of an embryonic stem cell such that that cell can be

distinguished from other cells. Exemplary distinguishing embryonic stem cell characteristics include, without limitation, gene expression profile, proliferative capacity, differentiation capacity, karyotype, responsiveness to particular culture conditions, and the like.

The term "adult stem cell" is used to refer to any multipotent stem cell derived from non-embryonic tissue, including fetal, juvenile, and adult tissue. Stem cells have been isolated from a wide variety of adult tissues including blood, bone marrow, brain, olfactory epithelium, skin, pancreas, skeletal muscle, and cardiac muscle. Each of these stem cells can be characterized based on gene expression, factor responsiveness, and morphology in culture.

"Proliferation" indicates an increase in cell number.

The term "tissue" refers to a group or layer of similarly specialized cells which together perform certain special functions.

The term "substantially pure", with respect to a particular cell population, refers to a population of cells that is at least about 75%, preferably at least about 85%, more preferably at least about 90%, and most preferably at least about 95% pure, with respect to the cells making up a total cell population. Recast, the term "substantially pure" refers to a population of cells that contain fewer than about 20%, more preferably fewer than about 10%, most preferably fewer than about 5%, of lineage committed cells.

The term "agents" includes one agent, more than one agent, or libraries of agents. By agents is meant to include nucleic acids, peptides, polypeptides, peptide mimetics, antisense oligonucleotides, RNAi constructs, antibodies, small organic molecules, and the like.

The invention further contemplates the screening of libraries of agents. Such libraries may include, without limitation, cDNA libraries (either plasmid based or phage based), expression libraries, combinatorial libraries, chemical libraries, phage display libraries, variegated libraries, and biased libraries. The term "library" refers to a collection of nucleic acids, proteins, peptides, chemical compounds, small organic molecules, or antibodies. Libraries comprising each of these are well known in the art. Exemplary types of libraries include combinatorial, variegated, biased, and unbiased libraries. Libraries can provide a systematic way to screen large

numbers of nucleic acids, proteins, peptides, peptide mimetics, chemical compounds, small organic molecules, or antibodies. Often, libraries are sub-divided into pools containing some fraction of the total species represented in the entire library. These pools can then be screened to identify fractions containing the desired activity. The pools can be further subdivided, and this process can be repeated until either (i) the desired activity can be correlated with a specific species contained within the library, or (ii) the desired activity is lost during further subdivision of the pool of species, and thus is the result of multiple species contained within the library.

A "marker" is used to determine the state of a cell. Markers are characteristics, whether morphological or biochemical (enzymatic), particular to a cell type, or molecules expressed by the cell type.

Markers may be detected by any method available to one of skill in the art. Exemplary methods of detecting nucleic acid markers include Northern blot, in situ hybridization, RT-PCR, RNAse protection, and the like. These methods can be coupled with a suitable detection system to allow for visualization of the result. Exemplary methods of detecting protein markers include immunohistochemistry (e.g., detecting ability of an antibody to bind to an epitope on a give protein of interest) and Western blot analysis. These methods can be coupled with a suitable detection system to allow for visualization of the result.

The term "reporter construct" is used to refer to constructs that 'report' or 'identify' the presence of particular cells. Typically reporter constructs include portions of the promoter, enhancer, or other regulatory sequences of a particular gene sufficient to regulate expression in a developmentally relevant manner. Such regulatory sequences are operably linked to a nucleic acid sequence encoding a marker that can be readily detectable (the 'reporter gene'). In this way, expression of a readily detectable product can be monitored, and this product is regulated in a manner consistent with the promoter or enhancer to which it is operably linked. Reporter genes may be introduced into cells by any of a number of ways including transfection, electroporation, micro-injection, etc. Exemplary reporter genes include, but are not limited to, green fluorescent protein (GFP), recombinantly engineered variants of GFP, red fluorescent protein, yellow fluorescent protein, cyan fluorescent

protein, LacZ, luciferase, firefly Remmila protein. Further exemplary reporter genes encode antibiotic resistance proteins including, but not limited to, neomycin, hygromycin, zeocine, and puromycin.

(iii) Exemplary Compositions

The present invention provides novel markers of embryonic stem cells. The novel embryonic stem cell markers are expressed in embryonic stem cells and germ cells, and are not expressed (or are expressed at a substantially decreased level) in embryoid bodies and other differentiated cell types. Accordingly, the embryonic stem cell markers provided herein provide novel compositions for identifying, purifying, and enriching populations of embryonic stem cells.

OCT4 was a previously identified marker of embryonic stem cells. The identification of OCT4 in the course of our search for novel ES-cell markers confirms the robustness of our studies which provide four novel ES cell markers. Nucleic acid sequences of the novel ES cell markers are provided in SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7, and SEQ ID NO: 9.

Given the identification of novel markers of embryonic stem cells, the present invention contemplates probes/primers for use in the identification and/or characterization and/or purification of embryonic stem cells. Exemplary probe/primers can be used to detect, in a cell or sample of cells, the cell or cells which express a nucleic acid represented in SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7, and SEQ ID NO: 9. Such probes/primers may themselves be labeled with a detectable label. Exemplary detectable labels include luciferase, radioactivity, a fluorescent moiety, and the like. Alternatively, detection may involve the use of a secondary reagent which is itself detectable. For example, detection of a nucleic acid by in situ hybridization often uses this approach. The probe which hybridizes to a specific sequence in a sample contains a labeled base. Often, digoxigenin is incorporated into the probe during synthesis. The dig labeled probes are later detected using an anti-dig antibody which is either itself labeled with a detectable moiety or which is further recognized by a secondary antibody with is labeled with a detectable moiety.

Exemplary probes/primers include a nucleic acid comprising a nucleic acid sequence that hybridizes under stringent conditions, including a wash step of 0.2X

SSC at 65 °C, to a nucleic acid represented in any of SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7, SEQ ID NO: 9, or the complement thereof. Further exemplary probes/primers include a nucleic acid comprising all or a portion of SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7, SEQ ID NO: 9, or the complement thereof. By a portion is meant to include fragments of the sequences represented in SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7, and SEQ ID NO: 9, wherein the fragments are at least 10, 12, 15, 18, 20, 22, 25, 50, 75, or greater than 75 nucleotides in length.

The invention further provides exemplary primer pairs which can be used to amplify (and thus detect) particular embryonic stem cell specific genes. Exemplary primer pairs for use in the manufacture of a composition for identifying embryonic stem cells, comprise the primer pair represented in SEQ ID NO: 11 and SEQ ID NO: 12, the primer pair represented in SEQ ID NO: 13 and SEQ ID NO: 14, the primer pair represented in SEQ ID NO: 15 and SEQ ID NO: 16, the primer pair represented in SEQ ID NO: 17 and SEQ ID NO: 18, and the primer pair represented in SEQ ID NO: 19 and SEQ ID NO: 20.

The foregoing probe/primers can be used in any of a number of methods as markers to detect the expression of an embryonic stem cell specific gene. Exemplary methods for detecting expression of a gene in a sample include Northern blot, in situ hybridization, RT-PCR, RNase protection, and micro-array analysis. These methods are routine in the art and are readily practiced. Methods for carrying out any of these methods, as well as methods for marking probes/primers appropriate for use in any one of these methods are routine in the art.

The invention further contemplates reporter constructs comprising a regulatory region (nucleic acid sequence) of any of the stem cell specific markers of the present invention operably linked to a reporter gene. The invention contemplates the reporter constructs, themselves, as well as cells engineered to express exogenously supplied reporter construct. Additionally, the invention contemplates the use of these reporter constructs and cells expressing the reporter constructs, in methods of identifying, characterizing, isolating, and purifying cells with a stem cell phenotype.

Exemplary reporter constructs contain all or a portion of the promoter or an enhancer of any of the stem cell specific markers of the invention (i.e., the markers

represented in SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7, or SEQ ID NO: 9). The portion of the promoter or enhancer must be sufficient to regulate expression of the operably linked reporter gene in a cell-specific manner (i.e., to regulate expression in cells with an embryonic stem cell or embryonic germ cell phenotype but not in cells which do not have an embryonic stem cell or germ cell phenotype.

Given the embryonic stem cell specific markers identified herein, one of skill in the art can readily identify portions of the promoters of these genes sufficient to regulate expression of reporter genes using standard methods of molecular biology. For example, one can use 5' –RACE to isolate the promoter of each of these genes. Fragments of the promoter can be sub-cloned and operably linked to a reporter gene (e.g., LacZ, GFP, luciferase, an antibiotic resistance gene), and these constructs can be tested in embryonic stem cells for proper, cell-type specific, regulated expression.

The invention additionally contemplates identification, isolation, and characterization of stem cells based on the expression of stem cell specific proteins. For example, characterization of cells based on expression of a polypeptide encoded by a stem cell specific nucleic acid comprising a nucleic acid sequence represented in SEQ ID NO: 1, 3, 5, 7, or 9, or a nucleic acid that hybridizes under stringent conditions to a nucleic acid represented in SEQ ID NO: 1, 3, 5, 7, or 9. By way of further example, the invention contemplates characterization of cells based on the expression of a polypeptide comprising all or a portion of an amino acid sequence represented in SEQ ID NO: 2, 4, 6, 8, or 10, or a variant thereof. By variant thereof, the invention contemplates characterization of cells based on the expression of a polypeptide comprising all or a portion of an amino acid sequence at least 80%, 90%, 95%, 98%, 99%, or greater than 99% identical to SEQ ID NO: 2, 4, 6, 8, or 10. By portion, the invention contemplates polypeptides of at least 12, 15, 20, 24, 30, 50, 75, or greater than 75 amino acid residues.

The invention further contemplates antibodies that are specifically immunoreactive with a protein encoded by any of the stem cell specific markers of the present invention, as polypeptide as the use of such antibodies in methods of identifying, characterizing, isolating, monitoring, and/or purifying cells. For example, the invention contemplates antibodies immunoreactive with a polypeptide encoded by a nucleic acid represented in SEQ ID NO: 1, 3, 5, 7, or 9, or by a nucleic

acid that hybridizes under stringent conditions to a nucleic acid represented in SEQ ID NO: 1, 3, 5, 7, or 9. Furthermore, the invention contemplates antibodies immunoreactive with a polypeptide comprising all or a portion of an amino acid sequence represented in SEQ ID NO: 2, 4, 6, 8, or 10, or a variant thereof at least 80%, 90%, 95%, 98%, 99%, or greater than 99% identical to all or a portion of an amino acid sequence represented in SEQ ID NO: 2, 4, 6, 8, or 10.

Antibodies can have extraordinary affinity and specificity for particular epitopes. Monoclonal or polyclonal antibodies can be made using standard protocols (See, for example, Antibodies: A Laboratory Manual ed. by Harlow and Lane (Cold Spring Harbor Press: 1988)). A mammal, such as a mouse, a hamster, a rat, a goat, or a rabbit can be immunized with an immunogenic form of the peptide. Techniques for conferring immunogenicity on a protein or peptide include conjugation to carriers or other techniques well known in the art.

Following immunization of an animal with an antigenic preparation of a polypeptide, antisera can be obtained and, if desired, polyclonal antibodies isolated from the serum. To produce monoclonal antibodies, antibody-producing cells (lymphocytes) can be harvested from an immunized animal and fused by standard somatic cell fusion procedures with immortalizing cells such as myeloma cells to yield hybridoma cells. Such techniques are well known in the art, and include, for example, the hybridoma technique (originally developed by Kohler and Milstein, (1975) Nature, 256: 495-497), the human B cell hybridoma technique (Kozbar et al., (1983) Immunology Today, 4: 72), and the EBV-hybridoma technique to produce human monoclonal antibodies (Cole et al., (1985) Monoclonal Antibodies, Alan R. Liss, Inc. pp. 77-96). Hybridoma cells can be screened immunochemically for production of antibodies specifically reactive with a particular polypeptide and monoclonal antibodies isolated from a culture comprising such hybridoma cells.

In the context of the present invention, antibodies can be screened and tested to identify those antibodies that are specifically immunoreactive with any of the embryonic stem cell specific markers of the present invention.

The term antibody as used herein is intended to include fragments thereof which are also specifically reactive with a particular polypeptide. Antibodies can be fragmented using conventional techniques and the fragments screened for utility in

the same manner as described above for whole antibodies. For example, $F(ab)_2$ fragments can be generated by treating antibody with pepsin. The resulting $F(ab)_2$ fragment can be treated to reduce disulfide bridges to produce Fab fragments. The antibody of the present invention is further intended to include bispecific and chimeric molecules having affinity for a particular protein conferred by at least one CDR region of the antibody.

Both monoclonal and polyclonal antibodies (Ab) directed against a particular polypeptides, and antibody fragments such as Fab, $F(ab)_2$, Fv and scFv can be used. The invention further contemplates the use of humanized antibodies which can readily made by one of skill in the art given a particular antibody generated in a non-human host.

(iv) Exemplary Methods

The systems and methods described herein also provide vectors containing an embryonic stem cell specific nucleic acid, operably linked to at least one transcriptional regulatory sequence. Such vectors may be used for expressing a polypeptide encoding an embryonic stem cell specific gene, delivering an embryonic stem cell specific gene to a cell, or in methods of making a probe/primer capable of detecting expression of said embryonic stem cell specific gene. Said probe/primer may optionally include a detectable label. The invention contemplates that certain vectors may be appropriate for both expressing a polypeptide encoded by a particular embryonic stem cell specific gene and for making a probe/primer which hybridizes under stringent conditions, including a wash step of 0.2X SSC at 65 °C, to an embryonic stem cell specific gene. The invention further contemplates, however, that certain vectors may be appropriate for preparation of a probe/primer but not for expressing a polypeptide encoded by a particular embryonic stem cells specific gene. One of skill in the art can readily select from amongst available vectors, as well as select whether the vector should include all or only a portion of a nucleic acid sequence corresponding to an embryonic stem cell specific gene.

Regulatory sequences are art-recognized and are selected to direct expression of the subject proteins. Accordingly, the term transcriptional regulatory sequence includes promoters, enhancers and other expression control elements. Such regulatory sequences are described in Goeddel; *Gene Expression Technology*:

Methods in Enzymology 185, Academic Press, San Diego, CA (1990). For instance, any of a wide variety of expression control sequences may be used in these vectors to express nucleic acid sequences encoding the agents of this invention. Such useful expression control sequences, include, for example, a viral LTR, such as the LTR of the Moloney murine leukemia virus, the LTR of the Herpes Simplex virus-1, the early and late promoters of SV40, adenovirus or cytomegalovirus immediate early promoter, the lac system, the trp system, the TAC or TRC system, T7 promoter whose expression is directed by T7 RNA polymerase, the major operator and promoter regions of phage λ, the control regions for fd coat protein, the promoter for 3-phosphoglycerate kinase or other glycolytic enzymes, the promoters of acid phosphatase, the promoters of the yeast α-mating factors, the polyhedron promoter of the baculovirus system and other sequences known to control the expression of genes of prokaryotic or eukaryotic cells or their viruses, and various combinations thereof. It should be understood that the design of the expression vector may depend on such factors as the choice of the host cell to be transformed and/or the type of protein desired to be expressed. Moreover, the vector's copy number, the ability to control that copy number and the expression of any other proteins encoded by the vector, such as antibiotic markers, should also be considered.

Moreover, the gene constructs can be used to deliver nucleic acids encoding the subject polypeptides. Thus, another aspect of the invention features expression vectors for *in vivo* or *in vitro* transfection, viral infection and expression of a subject polypeptide in particular cell types.

This application also describes methods for producing the subject polypeptides. For example, a host cell transfected with a nucleic acid vector directing expression of a nucleotide sequence encoding the subject polypeptides can be cultured under appropriate conditions to allow expression of the peptide to occur. The polypeptide may be secreted and isolated from a mixture of cells and medium containing the recombinant polypeptide. Alternatively, the peptide may be expressed cytoplasmically and the cells harvested, lysed and the protein isolated. A cell culture includes host cells, media and other by-products. Suitable media for cell culture are well known in the art. The recombinant polypeptide can be isolated from cell culture medium, host cells, or both using techniques known in the art for

purifying proteins including ion-exchange chromatography, gel filtration chromatography, ultrafiltration, electrophoresis, and immunoaffinity purification with antibodies specific for such peptide. In one example, the recombinant polypeptide is a fusion protein containing a domain which facilitates its purification, such as a GST fusion protein. In another example, the subject recombinant polypeptide may include one or more additional domains which facilitate immunodetection, purification, and the like. Exemplary domains include HA, FLAG, GST, His, and the like. Further exemplary domains include a protein transduction domain (PTD) which facilitates the uptake of proteins by cells.

This application also describes a host cell which expresses a recombinant form of the subject polypeptides. The host cell may be a prokaryotic or eukaryotic cell. Thus, a nucleotide sequence derived from the cloning of a protein encoding all or a selected portion (either an antagonistic portion or a bioactive fragment) of the full-length protein, can be used to produce a recombinant form of a polypeptide via microbial or eukaryotic cellular processes. Ligating the polynucleotide sequence into a gene construct, such as an expression vector, and transforming or transfecting into hosts, either eukaryotic (yeast, avian, insect or mammalian) or prokaryotic (bacterial cells), are standard procedures used in producing other well-known proteins, e.g. insulin, interferons, human growth hormone, IL-1, IL-2, and the like. Similar procedures, or modifications thereof, can be employed to prepare recombinant polypeptides by microbial means or tissue-culture technology in accord with the subject invention. Such methods are used to produce experimentally useful proteins that include all or a portion of the subject nucleic acids. For example, such methods are used to produce fusion proteins including domains which facilitate purification or immunodetection, and to produce recombinant forms of a protein.

The recombinant genes can be produced by ligating a nucleic acid encoding a protein, or a portion thereof, into a vector suitable for expression in either prokaryotic cells, eukaryotic cells, or both. Expression vectors for production of recombinant forms of the subject polypeptides include plasmids and other vectors. For instance, suitable vectors for the expression of a polypeptide include plasmids of the types: pBR322-derived plasmids, pEMBL-derived plasmids, pEX-derived plasmids, pGEX-derived plasmids, pTrc-His-derived plasmids, pBTac-derived

plasmids and pUC-derived plasmids for expression in prokaryotic cells, such as *E. coli*. Similarly, appropriate vectors can be readily selected for maintaining embryonic stem cell specific nucleic acid sequences, and for preparing a probe/primer that hybridizes under stringent conditions, including a wash step of 0.2X SSC at 65 °C, to an embryonic stem cell specific nucleic acid. In one embodiment, the probe/primer hybridizes under stringent conditions, including a wash step of 0.2X SSC at 65 °C, to a nucleic acid sequence represented in any of SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7, or SEQ ID NO: 9.

A number of vectors exist for the expression of recombinant proteins in yeast. For instance, YEP24, YIP5, YEP51, YEP52, pYES2, and YRP17 are cloning and expression vehicles useful in the introduction of genetic constructs into *S. cerevisiae*.

Many mammalian expression vectors contain both prokaryotic sequences, to facilitate the propagation of the vector in bacteria, and one or more eukaryotic transcription units that are expressed in eukaryotic cells. The pcDNAI/amp, pcDNAI/neo, pRc/CMV, pSV2gpt, pSV2neo, pSV2-dhfr, pTk2, pRSVneo, pMSG, pSVT7, pko-neo, pBacMam-2, and pHyg derived vectors are examples of mammalian expression vectors suitable for transfection of eukaryotic cells. Some of these vectors are modified with sequences from bacterial plasmids, such as pBR322, to facilitate replication and drug resistance selection in both prokaryotic and eukaryotic cells. For other suitable expression systems for both prokaryotic and eukaryotic cells, as well as general recombinant procedures, see *Molecular Cloning A Laboratory Manual*, 3rd Ed., ed. by Sambrook and Russell (Cold Spring Harbor Laboratory Press: 2001).

In some instances, it may be desirable to express the recombinant polypeptides by the use of a baculovirus expression system. Examples of such baculovirus expression systems include pVL-derived vectors (such as pVL1392, pVL1393 and pVL941), pAcUW-derived vectors (such as pAcUW1), and pBlueBac-derived vectors (such as the ß-gal containing pBlueBac III).

When it is desirable to express only a portion of a protein, such as a form lacking a portion of the N-terminus, e.g. a truncation mutant, it may be necessary to add a start codon (ATG) to the oligonucleotide fragment containing the desired sequence to be expressed. It is well known in the art that a methionine at the N-

terminal position can be enzymatically cleaved by the enzyme methionine aminopeptidase (MAP).

Techniques for making fusion genes are known to those skilled in the art. The joining of various nucleic acid fragments coding for different polypeptide sequences is performed in accordance with conventional techniques, employing blunt-ended or stagger-ended termini for ligation, restriction enzyme digestion to provide for appropriate termini, filling-in of cohesive ends as appropriate, alkaline phosphatase treatment to avoid undesirable joining, and enzymatic ligation. In another example, the fusion gene can be synthesized by conventional techniques including automated DNA synthesizers. Alternatively, PCR amplification of gene fragments can be carried out using anchor primers which give rise to complementary overhangs between two consecutive gene fragments which can subsequently be annealed to generate a chimeric gene sequence.

The present invention also makes available isolated polypeptides which are isolated from, or otherwise substantially free of other cellular and extracellular proteins. The term "substantially free of other cellular or extracellular proteins" (also referred to herein as "contaminating proteins") or "substantially pure or purified preparations" are defined as encompassing preparations having less than 20% (by dry weight) contaminating protein, and preferably having less than 5% contaminating protein. Functional forms of the subject polypeptides can be prepared as purified preparations by using a cloned gene as described herein. By "purified", it is meant, when referring to peptide or nucleic acid sequences, that the indicated molecule is present in the substantial absence of other biological macromolecules, such as other proteins. The term "purified" as used herein preferably means at least 80% by dry weight, more preferably in the range of 95-99% by weight, and most preferably at least 99.8% by weight, of biological macromolecules of the same type present (but water and buffers can be present). The term "pure" as used herein preferably has the same numerical limits as "purified" immediately above. "Isolated" and "purified" do not encompass either natural materials in their native state or natural materials that have been separated into components (e.g., in an acrylamide gel) but not obtained either as pure (e.g. lacking contaminating proteins,

or chromatography reagents such as denaturing agents and polymers, e.g. acrylamide or agarose) substances or solutions.

Isolated peptidyl portions of proteins can be obtained by screening peptides recombinantly produced from the corresponding fragment of the nucleic acid encoding such peptides. In addition, fragments can be chemically synthesized using techniques known in the art such as conventional Merrifield solid phase f-Moc or t-Boc chemistry. The recombinant polypeptides of the present invention also include versions of those proteins that are resistant to proteolytic cleavage. Variants of the present invention also include proteins which have been post-translationally modified in a manner different than the authentic protein. Modification of the structure of the subject polypeptides can be for such purposes as enhancing therapeutic or prophylactic efficacy, or stability (e.g., *ex vivo* shelf life and resistance to proteolytic degradation *in vivo*).

For example, it is reasonable to expect that, in some instances, an isolated replacement of a leucine with an isoleucine or valine, an aspartate with a glutamate, a threonine with a serine, or a similar replacement of an amino acid with a structurally related amino acid (e.g., isosteric and/or isoelectric mutations) may not have a major effect on the biological activity of the resulting molecule. Conservative replacements are those that take place within a family of amino acids that are related in their side chains. Genetically encoded amino acids can be divided into four families: (1) acidic = aspartate, glutamate; (2) basic = lysine, arginine, histidine; (3) nonpolar = alanine, valine, leucine, isoleucine, proline, phenylalanine, methionine, tryptophan; and (4) uncharged polar = glycine, asparagine, glutamine, cysteine, serine, threonine, tyrosine. Phenylalanine, tryptophan, and tyrosine are sometimes classified jointly as aromatic amino acids. In similar fashion, the amino acid repertoire can be grouped as (1) acidic = aspartate, glutamate; (2) basic = lysine, arginine histidine, (3) aliphatic = glycine, alanine, valine, leucine, isoleucine, serine, threonine, with serine and threonine optionally be grouped separately as aliphatic-hydroxyl; (4) aromatic = phenylalanine, tyrosine, tryptophan; (5) amide = asparagine, glutamine; and (6) sulfur -containing = cysteine and methionine. (see, for example, *Biochemistry*, 5th ed. by Berg, Tymoczko and Stryer, WH Freeman and Co.: 2002). Whether a change in the amino acid sequence of a peptide results in a variant which

maintains the same function as the wildtype protein, or a variant which antagonizes the function of the wildtype protein, can be determined by assessing the ability of the variant peptide to produce a response in cells in a fashion similar to the wild-type protein, or antagonize such a response. Polypeptides in which more than one replacement has taken place can readily be tested in the same manner.

Advances in the fields of combinatorial chemistry and combinatorial mutagenesis have facilitated the making of polypeptide variants (Wissmanm et al. (1991) *Genetics* **128**: 225-232; Graham et al. (1993) *Biochemistry* **32**: 6250-6258; York et al. (1991) *Journal of Biological Chemistry* **266**: 8495-8500; Reidhaar-Olson et al. (1988) *Science* **241**: 53-57). Given one or more assays for testing polypeptide variants, one can assess whether a given variant functions as an antagonist, or whether a given variant has the same or substantially the same function as the wildtype protein. In the context of the present invention, several methods for assaying the functional activity of potential variants are provided.

To further illustrate, the invention contemplates a method for generating sets of combinatorial mutants, as well as truncation mutants, and is especially useful for identifying potential agonistic or antagonistic variant sequences. The purpose of screening such combinatorial libraries is to generate, for example, novel variants which can agonize or antagonize the function of an embryonic stem cell specific gene. Such variants may be useful to either promote an embryonic stem cell phenotype or to inhibit an embryonic stem cell phenotype.

However, in addition to the use of agonistic or antagonistic polypeptide variant sequences, the invention contemplates the use of agonistic or antagonistic nucleic acids comprising the embryonic stem cell specific nucleic acid sequences provided in SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7, or SEQ ID NO: 9 in the preparation of compositions to modulate an embryonic stem cell phenotype in a cell.

For example, the invention contemplates the use of antisense oligonucleotides, ribozymes, and RNAi constructs capable of inhibiting the expression and/or activity of any of the stem cell specific markers of the present invention. In one embodiment, the antisense oligonucleotide, ribozyme, or RNAi construct comprises a nucleic acid capable of hybridizing under stringent conditions,

including a wash step of 0.2X SSC at 65 °C, to a nucleic acid sequence represented in SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7, or SEQ ID NO: 9.

In one example, a variegated library of variants is generated by combinatorial mutagenesis at the nucleic acid level, and is encoded by a variegated gene library. For instance, a mixture of synthetic oligonucleotides can be enzymatically ligated into gene sequences such that the degenerate set of potential sequences are expressible as individual polypeptides, or alternatively, as a set of larger fusion proteins (e.g. for phage display) containing the set of sequences therein.

The library of potential variants can be generated from a degenerate oligonucleotide sequence using a variety of methods. Chemical synthesis of a degenerate gene sequence can be carried out in an automatic DNA synthesizer, and the synthetic genes then ligated into an appropriate expression vector. One purpose of a degenerate set of genes is to provide, in one mixture, all the sequences encoding the desired set of potential variant sequences. The synthesis of degenerate oligonucleotides is known in the art.

A range of techniques are known for screening gene products of combinatorial libraries made by point mutations, and for screening cDNA libraries for gene products having a certain property. Such techniques will be generally adaptable for rapid screening of the gene libraries generated by combinatorial mutagenesis. These techniques are also applicable for rapid screening of other gene libraries. One example of the techniques used for screening large gene libraries includes cloning the gene library into replicable expression vectors, transforming appropriate cells with the resulting library of vectors, and expressing the combinatorial genes under conditions in which detection of a desired activity facilitates relatively easy isolation of the vector encoding the gene whose product was detected.

Constructs comprising the subject agents may be administered in biologically effective carriers, e.g. any formulation or composition capable of effectively delivering the agents to cells *in vivo* or *in vitro*. The particular approach can be selected from amongst those well known to one of skill in the art based on the particular agent to be delivered (e.g., DNA enzyme, polypeptide variant, peptidomimetic, RNAi construct, antibody, antisense oligonucleotide, small organic

molecule, and the like), the cell type to which delivery is desired, and the route of administration.

Approaches include viral vectors including recombinant retroviruses, adenovirus, adeno-associated virus, herpes simplex virus-1, lentivirus, mammalian baculovirus or recombinant bacterial or eukaryotic plasmids. Viral vectors transfect cells directly; plasmid DNA can be delivered with the help of, for example, cationic liposomes (lipofectin) or derivatized (e.g. antibody conjugated), polylysine conjugates, gramacidin S, artificial viral envelopes or other such intracellular carriers, as well as direct injection of the gene construct, electroporation or $CaPO_4$ precipitation. One of skill in the art can readily select from available vectors and methods of delivery in order to optimize expression in a particular cell type or under particular conditions.

Retrovirus vectors and adeno-associated virus vectors have been frequently used for the transfer of exogenous genes. These vectors can be used to deliver nucleic acids, for example RNAi constructs, as well as to deliver nucleic acids encoding particular proteins such as polypeptide variants. These vectors provide efficient delivery of genes into cells. A major prerequisite for the use of retroviruses is to ensure the safety of their use, particularly with regard to the possibility of the spread of wild-type virus in the cell population. The development of specialized cell lines (termed "packaging cells") which produce only replication-defective retroviruses has increased the utility of retroviruses for gene therapy, and defective retroviruses are well characterized for use in gene transfer for gene therapy purposes. Thus, recombinant retrovirus can be constructed in which part of the retroviral coding sequence (*gag, pol, env*) has been replaced by nucleic acid encoding one of the subject proteins rendering the retrovirus replication defective. The replication defective retrovirus is then packaged into virions through the use of a helper virus by standard techniques which can be used to infect a target cell. Protocols for producing recombinant retroviruses and for infecting cells *in vitro* or *in vivo* with such viruses can be found in Current Protocols in Molecular Biology, Ausubel, F.M. et al. (eds.) Greene Publishing Associates, (2000), and other standard laboratory manuals. Examples of suitable retroviruses include pBPSTR1, pLJ, pZIP, pWE and pEM which are known to those skilled in the art. Examples of suitable packaging

virus lines for preparing both ecotropic and amphotropic retroviral systems include ψCrip, ψCre, ψ2, ψAm, and PA317.

Furthermore, it has been shown that it is possible to limit the infection spectrum of retroviruses and consequently of retroviral-based vectors, by modifying the viral packaging proteins on the surface of the viral particle (see, for example PCT publications WO93/25234 and WO94/06920). For instance, strategies for the modification of the infection spectrum of retroviral vectors include: coupling antibodies specific for cell surface antigens to the viral *env* protein; or coupling cell surface receptor ligands to the viral *env* proteins. Coupling can be in the form of the chemical cross-linking with a protein or other variety (e.g. lactose to convert the *env* protein to an asialoglycoprotein), as well as by generating fusion proteins (e.g. single-chain antibody/*env* fusion proteins). This technique, while useful to limit or otherwise direct the infection to certain tissue types, can also be used to convert an ecotropic vector into an amphotropic vector.

Moreover, use of retroviral gene delivery can be further enhanced by the use of tissue- or cell-specific transcriptional regulatory sequences which control expression of the gene of the retroviral vector such as tetracycline repression or activation.

Another viral gene delivery system which has been employed utilizes adenovirus-derived vectors. The genome of an adenovirus can be manipulated so that it encodes and expresses a gene product of interest but is inactivated in terms of its ability to replicate in a normal lytic viral life cycle. Suitable adenoviral vectors derived from the adenovirus strain Ad type 5 dl324 or other strains of adenovirus (e.g., Ad2, Ad3, Ad7 etc.) are known to those skilled in the art. Recombinant adenoviruses can be advantageous in certain circumstances in that they can be used to infect a wide variety of cell types, including airway epithelium, endothelial cells, hepatocytes, and muscle cells. Furthermore, the virus particle is relatively stable and amenable to purification and concentration, and as above, can be modified so as to affect the spectrum of infectivity.

Yet another viral vector system is the adeno-associated virus (AAV). Adeno-associated virus is a naturally occurring defective virus that requires another virus, such as an adenovirus or a herpes virus, as a helper virus for efficient

replication and a productive life cycle. (For a review see Muzyczka et al. *Curr. Topics in Micro. and Immunol.* (1992) **158**: 97-129). It is also one of the few viruses that may integrate its DNA into non-dividing cells, and exhibits a high frequency of stable integration.

Another viral delivery system is based on herpes simplex-1 (HSV-1). HSV-1 based vectors may be especially useful in the methods of the present invention because they have been previously shown to infect neuronal cells. Given that many adult neuronal cells are post-mitotic, and thus have been difficult to infect using some other commonly employed viruses, the use of HSV-1 represents a substantial advance and further underscores the potential utility of viral based systems to facilitate gene expression in the nervous system (Agudo et al. (2002) *Human Gene Therapy* **13**: 665-674; Latchman (2001) *Neuroscientist* **7**: 528-537; Goss et al. (2002) *Diabetes* **51**: 2227-2232; Glorioso (2002) *Current Opin Drug Discov Devel* **5**: 289-295; Evans (2002) *Clin Infect Dis* **35**: 597-605; Whitley (2002) *Journal of Clinical Invest* **110**: 145-151; Lilley (2001) *Curr Gene Ther* **1**: 339-359).

The above cited examples of viral vectors are by no means exhaustive. However, they are provided to indicate that one of skill in the art may select from well known viral vectors, and select a suitable vector for expressing a particular protein in a particular cell type.

In addition to viral transfer methods, such as those illustrated above, non-viral methods can be used. Many nonviral methods of gene transfer rely on normal mechanisms used by cells for the uptake and intracellular transport of macromolecules. Exemplary gene delivery systems of this type include liposomal derived systems, poly-lysine conjugates, and artificial viral envelopes.

It may sometimes be desirable to introduce a nucleic acid directly to a cell, for example a cell in culture or a cell in an animal. Such administration can be done by injection of the nucleic acid (e.g., DNA, RNA) directly at the desired site. Such methods are commonly used in the vaccine field, specifically for administration of "DNA vaccines", and include condensed DNA (US Patent No. 6,281,005).

In addition to administration of nucleic acids, the systems and methods described herein contemplate that polypeptides may be administered directly. Some proteins, for example factors that act extracellularly by contacting a cell surface

receptor, such as growth factors, may be administered by simply contacting cells with said protein. For example, cells are typically cultured in media which is supplemented by a number of proteins such as FGF, TGFβ, insulin, etc. These proteins influence cells by simply contacting the cells. Such a method similarly pertains to other agents such as small organic molecules and chemical compounds. These agents may either exert their effect at the cell surface, or may be able to permeate the cell membrane without the need for additional manipulation.

In another embodiment, a polypeptide is directly introduced into a cell. Methods of directly introducing a polypeptide into a cell include, but are not limited to, protein transduction and protein therapy. For example, a protein transduction domain (PTD) can be fused to a nucleic acid encoding a particular polypeptide, and the fusion protein is expressed and purified. Fusion proteins containing the PTD are permeable to the cell membrane, and thus cells can be directly contacted with a fusion protein (Derossi et al. (1994) *Journal of Biological Chemistry* **269**: 10444-10450; Han et al. (2000) *Molecules and Cells* **6**: 728-732; Hall et al. (1996) *Current Biology* 6: 580-587; Theodore et al. (1995) *Journal of Neuroscience* **15**: 7158-7167).

Although some protein transduction based methods rely on fusion of a polypeptide of interest to a sequence which mediates introduction of the protein into a cell, other protein transduction methods do not require covalent linkage of a protein of interest to a transduction domain. At least two commercially available reagents exist that mediate protein transduction without covalent modification of the protein (Chariot™, produced by Active Motif, www.activemotif.com and Bioporter® Protein Delivery Reagent, produced by Gene Therapy Systems, www.genetherapysystems.com).

Briefly, these protein transduction reagents can be used to deliver proteins, peptides and antibodies directly to cells including mammalian cells. Delivery of proteins directly to cells has a number of advantages. Firstly, many current techniques of gene delivery are based on delivery of a nucleic acid sequence which must be transcribed and/or translated by a cell before expression of the protein is achieved. This results in a time lag between delivery of the nucleic acid and expression of the protein. Direct delivery of a protein decreases this delay.

Secondly, delivery of a protein often results in transient expression of the protein in a cell.

As outlined herein, protein transduction mediated by covalent attachment of a PTD to a protein can be used to deliver a protein to a cell. These methods require that individual proteins be covalently appended with PTD moieties. In contrast, methods such as Chariot™ and Bioporter® facilitate transduction by forming a noncovalent interaction between the reagent and the protein. Without being bound by theory, these reagents are thought to facilitate transit of the cell membrane, and following internalization into a cell the reagent and protein complex disassociates so that the protein is free to function in the cell.

(v) Methods of administration of nucleic acids, proteins, chemical compounds and pharmaceutical compositions of agents

Probes/primers and agents for use in the methods of the present invention, as well as agents identified by the subject methods may be conveniently formulated for administration with a biologically acceptable medium, such as water, buffered saline, polyol (for example, glycerol, propylene glycol, liquid polyethylene glycol and the like) or suitable mixtures thereof. Optimal concentrations of the active ingredient(s) in the chosen medium can be determined empirically, according to procedures well known to medicinal chemists. As used herein, "biologically acceptable medium" includes solvents, dispersion media, and the like which may be appropriate for the desired route of administration of the one or more agents. The use of media for pharmaceutically active substances is known in the art. Except insofar as a conventional media or agent is incompatible with the activity of a particular agent or combination of agents, its use in the pharmaceutical preparation of the invention is contemplated. Suitable vehicles and their formulation inclusive of other proteins are described, for example, in the book *Remington's Pharmaceutical Sciences* (Remington's Pharmaceutical Sciences. Mack Publishing Company, Easton, Pa., USA 1985). These vehicles include injectable "deposit formulations".

(vi) Screening Assays

The present invention provides unique markers of embryonic stem cells. Given that the expression of these markers is consistent with the embryonic stem cell phenotype, it is contemplated that agents that increase the expression of any of

these markers can also be used to promote the embryonic stem cell phenotype of a cell. Accordingly, the present invention contemplates screening to identify and/or characterize agents that promote the embryonic stem cell phenotype of a cell.

Similarly, the present invention contemplates the identification and/or characterization of agents that inhibit the expression of one or more of the stem cell specific markers provided herein. Such agents can be used to inhibit the embryonic stem cell phenotype of a cell.

Agents for use in the methods of the present invention include nucleic acids, peptides, polypeptides, peptide mimetics, small organic molecules, antisense oligonucleotides, RNAi constructs, ribozymes, antibodies, and the like. Agents can be screened individually, in combination, or as a library of agents.

An exemplary method of identifying agents that promote the embryonic stem cell phenotype is as follows. Briefly, the invention contemplates contacting one or more cells, wherein the cell is not an embryonic stem cell, with one or more agents, and identifying the agents that increase the expression of at least one embryonic stem cell marker (i.e., SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7, or SEQ ID NO: 9). Agents that increase the expression of any one of SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7 or SEQ ID NO: 9 can then be further analyzed to assess further effects on the phenotype of the cell. For example, a cell contacted with the candidate agent can be assayed for expression of the known ES cell marker OCT4. Furthermore, such cells can be analyzed for other stem cell characteristics.

The invention contemplates the identification of (i) agents that promote the expression of any one of SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7, or SEQ ID NO: 9, (ii) agents that promote the expression of more than one marker selected from SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7, or SEQ ID NO: 9, (iii) agents that promote the expression of at least one of SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7, or SEQ ID NO: 9, and which promote one or more other characteristic of an embryonic stem cell phenotype.

As mentioned above, the present invention further contemplates the identification and/or characterization of agents that inhibit the expression of one or more of the stem cell specific markers provided herein. Such agents can be used to inhibit the embryonic stem cell phenotype of a cell.

Agents for use in the method of the present invention include nucleic acids, peptides, polypeptides, peptide mimetics, small organic molecules, antisense oligonucleotides, RNAi constructs, ribozymes, antibodies, and the like. Agents can be screened individually, in combination, or as a library of agents.

An exemplary method of identifying agents that inhibit the embryonic stem cell phenotype is as follows. Briefly, the invention contemplates contacting one or more embryonic stem cells, wherein the cells express the stem cell specific markers represented in SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7, and SEQ ID NO: 9, with one or more agents, and identifying the agents that decrease the expression of at least one embryonic stem cell marker (i.e., SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7, or SEQ ID NO: 9). Agents that decrease the expression of any one of SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7 or SEQ ID NO: 9 can then be further analyzed to evaluate whether the agent inhibits the embryonic stem cell phenotype of the cell. For example, a cell contacted with the candidate agent can be assayed for expression of the known ES cell marker OCT4. Furthermore, such cells can be analyzed for other stem cell characteristics.

The invention contemplates the identification of (i) agents that inhibit the expression of any one of SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7, or SEQ ID NO: 9, (ii) agents that inhibit the expression of more than one marker selected from SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7, or SEQ ID NO: 9, (iii) agents that inhibit the expression of at least one of SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7, or SEQ ID NO: 9, and which inhibit one or more other characteristic of an embryonic stem cell phenotype.

Similarly, the invention contemplates methods of identifying agents that promote or inhibit characteristics of embryonic stem cells based on examining expression of a polypeptide encoded by a nucleic acid comprising SEQ ID NO: 3, 5, 7, 9, or a nucleic acid that hybridizes under stringent conditions with SEQ ID NO: 3, 5, 7, or 9. Such polypeptides include polypeptide comprising all or a portion of the amino acid sequence represented in SEQ ID NO: 4, 6, 8, or 10.

Agents identified by the foregoing methods, including both agents that promote and agents that inhibit the expression of an embryonic stem cell marker, can be formulated as a pharmaceutical preparation. Such a pharmaceutical

preparation comprises said agent and a pharmaceutically acceptable carrier or excipient.

As outlined in detail above, the invention contemplates that numerous classes of agents can be used to promote or inhibit the expression and/or activity of an embryonic stem cell specific marker. Exemplary agents include, but are not limited to nucleic acids, peptides, polypeptides, peptide mimetics, antisense oligonucleotides, RNAi constructs, ribozymes, small organic molecules, antibodies, and the like. The invention contemplates screening to identify such agents, and the invention further contemplates methods of promoting or inhibiting the embryonic stem cell phenotype (e.g., promoting the progress of stem cell differentiation) of a cell by inhibiting expression and/or activity of an embryonic stem cell specific marker.

Numerous mechanisms exist to inhibit the expression and/or activity of a particular mRNA or protein. Without being bound by theory, the present invention contemplates any of a number of methods for inhibiting the expression and/or activity of a particular mRNA. Furthermore, the invention contemplates combinatorial methods comprising the use of two or more inhibitors that decrease the expression and/or activity of a particular mRNA or protein.

The following are illustrative examples of methods for inhibiting the expression and/or activity of an mRNA or protein. These examples are in no way meant to be limiting, and one of skill in the art can readily select from among known methods of inhibiting expression and/or activity.

Antisense oligonucleotides are relatively short nucleic acids that are complementary (or antisense) to the coding strand (sense strand) of the mRNA encoding a particular protein. Although antisense oligonucleotides are typically RNA based, they can also be DNA based. Additionally, antisense oligonucleotides are often modified to increase their stability.

Without being bound by theory, the binding of these relatively short oligonucleotides to the mRNA is believed to induce stretches of double stranded RNA that trigger degradation of the messages by endogenous RNAses. Additionally, sometimes the oligonucleotides are specifically designed to bind near the promoter of the message, and under these circumstances, the antisense

oligonucleotides may additionally interfere with translation of the message. Regardless of the specific mechanism by which antisense oligonucleotides function, their administration to a cell or tissue allows the degradation of the mRNA encoding a specific protein. Accordingly, antisense oligonucleotides decrease the expression and/or activity of a particular protein.

The oligonucleotides can be DNA or RNA or chimeric mixtures or derivatives or modified versions thereof, single-stranded or double-stranded. The oligonucleotide can be modified at the base moiety, sugar moiety, or phosphate backbone, for example, to improve stability of the molecule, hybridization, etc. The oligonucleotide may include other appended groups such as peptides (e.g., for targeting host cell receptors), or agents facilitating transport across the cell membrane (see, e.g., Letsinger et al., 1989, Proc. Natl. Acad. Sci. U.S.A. 86:6553-6556; Lemaitre et al., 1987, Proc. Natl. Acad. Sci. 84:648-652; PCT Publication No. W088/09810, published December 15, 1988) or the blood- brain barrier (see, e.g., PCT Publication No. W089/10134, published April 25, 1988), hybridization-triggered cleavage agents (See, e.g., Krol et al., 1988, BioTechniques 6:958- 976) or intercalating agents. (See, e.g., Zon, 1988, Pharm. Res. 5:539-549). To this end, the oligonucleotide may be conjugated to another molecule.

The antisense oligonucleotide may comprise at least one modified base moiety which is selected from the group including but not limited to 5-fluorouracil, 5- bromouracil, 5-chlorouracil, 5-iodouracil, hypoxanthine, xanthine, 4-acetylcytosine, 5- (carboxyhydroxytriethyl) uracil, 5-carboxymethylaminomethyl-2-thiouridine, 5- carboxymethylaminomethyluracil, dihydrouracil, beta-D-galactosylqueosine, inosine, N6- isopentenyladenine, 1-methylguanine, 1-methylinosine, 2,2-dimethylguanine, 2-methyladenine, 2-methylguanine, 3-methylcytosine, 5-methylcytosine, N6-adenine, 7-methylguanine, 5-methylaminomethyluracil, 5-methoxyaminomethyl-2-thiouracil, beta-D-mannosylqueosine, 5'-methoxycarboxymethyluracil, 5-methoxyuracil, 2-methylthio-N6- isopentenyladenine, uracil-5-oxyacetic acid (v), wybutoxosine, pseudouracil, queosine, 2-thiocytosine, 5-methyl-2-thiouracil, 2-thiouracil, 4-thiouracil, 5-methyluracil, uracil-5- oxyacetic acid methyl ester, uracil-5-oxyacetic acid (v), 5-

methyl-2-thiouracil, 3-(3-amino-3- N-2-carboxypropyl) uracil, (acp3)w, and 2,6-diaminopurine.

The antisense oligonucleotide may also comprise at least one modified sugar moiety selected from the group including but not limited to arabinose, 2-fluoroarabinose, xylulose, and hexose.

The antisense oligonucleotide can also contain a neutral peptide-like backbone. Such molecules are termed peptide nucleic acid (PNA)-oligomers and are described, e.g., in Perry-O'Keefe et al. (1996) Proc. Natl. Acad. Sci. U.S.A. 93:14670 and in Eglom et al. (1993) Nature 365:566. One advantage of PNA oligomers is their capability to bind to complementary DNA essentially independently from the ionic strength of the medium due to the neutral backbone of the DNA. In yet another embodiment, the antisense oligonucleotide comprises at least one modified phosphate backbone selected from the group consisting of a phosphorothioate, a phosphorodithioate, a phosphoramidothioate, a phosphoramidate, a phosphordiamidate, a methylphosphonate, an alkyl phosphotriester, and a formacetal or analog thereof.

In yet a further embodiment, the antisense oligonucleotide is an -anomeric oligonucleotide. An -anomeric oligonucleotide forms specific double-stranded hybrids with complementary RNA in which, contrary to the usual -units, the strands run parallel to each other (Gautier et al., 1987, Nucl. Acids Res. 15:6625-6641). The oligonucleotide is a 2'-0-methylribonucleotide (Inoue et al., 1987, Nucl. Acids Res. 15:6131-6148), or a chimeric RNA-DNA analogue (Inoue et al., 1987, FEBS Lett. 215:327-330).

Oligonucleotides of the invention may be synthesized by standard methods known in the art, e.g., by use of an automated DNA synthesizer (such as are commercially available from Biosearch, Applied Biosystems, etc.). As examples, phosphorothioate oligonucleotides may be synthesized by the method of Stein et al. (1988, Nucl. Acids Res. 16:3209), methylphosphonate oligonucleotides can be prepared by use of controlled pore glass polymer supports (Sarin et al., 1988, Proc. Natl. Acad. Sci. U.S.A. 85:7448-7451), etc.

The selection of an appropriate oligonucleotide can be readily performed by

one of skill in the art. Given the nucleic acid sequence encoding a particular protein, one of skill in the art can design antisense oligonucleotides that bind to that protein, and test these oligonucleotides in an in vitro or in vivo system to confirm that they bind to and mediate the degradation of the mRNA encoding the particular protein. To design an antisense oligonucleotide that specifically binds to and mediates the degradation of a particular protein, it is important that the sequence recognized by the oligonucleotide is unique or substantially unique to that particular protein. For example, sequences that are frequently repeated across protein may not be an ideal choice for the design of an oligonucleotide that specifically recognizes and degrades a particular message. One of skill in the art can design an oligonucleotide, and compare the sequence of that oligonucleotide to nucleic acid sequences that are deposited in publicly available databases to confirm that the sequence is specific or substantially specific for a particular protein.

In another example, it may be desirable to design an antisense oligonucleotide that binds to and mediates the degradation of more than one message. In one example, the messages may encode related protein such as isoforms or functionally redundant protein. In such a case, one of skill in the art can align the nucleic acid sequences that encode these related proteins, and design an oligonucleotide that recognizes both messages.

A number of methods have been developed for delivering antisense DNA or RNA to cells; e.g., antisense molecules can be injected directly into the tissue site, or modified antisense molecules, designed to target the desired cells (e.g., antisense linked to peptides or antibodies that specifically bind receptors or antigens expressed on the target cell surface) can be administered systematically.

However, it may be difficult to achieve intracellular concentrations of the antisense sufficient to suppress translation on endogenous mRNAs in certain instances. Therefore another approach utilizes a recombinant DNA construct in which the antisense oligonucleotide is placed under the control of a strong pol III or pol II promoter. For example, a vector can be introduced in vivo such that it is taken up by a cell and directs the transcription of an antisense RNA. Such a vector can remain episomal or become chromosomally integrated, as long as it can be transcribed to produce the desired antisense RNA. Such vectors can be constructed

by recombinant DNA technology methods standard in the art. Vectors can be plasmid, viral, or others known in the art, used for replication and expression in mammalian cells. Expression of the sequence encoding the antisense RNA can be by any promoter known in the art to act in mammalian, preferably human cells. Such promoters can be inducible or constitutive. Such promoters include but are not limited to: the SV40 early promoter region (Bernoist and Chambon, 1981, Nature 290:304-310), the promoter contained in the 3' long terminal repeat of Rous sarcoma virus (Yamamoto et al., 1980, Cell 22:787-797), the herpes thymidine kinase promoter (Wagner et al., 1981, Proc. Natl. Acad. Sci. U.S.A. 78:1441-1445), the regulatory sequences of the metallothionein gene (Brinster et al, 1982, Nature 296:39-42), etc. Any type of plasmid, cosmid, YAC or viral vector can be used to prepare the recombinant DNA construct that can be introduced directly into the tissue site. Alternatively, viral vectors can be used which selectively infect the desired tissue, in which case administration may be accomplished by another route (e.g., systematically).

RNAi constructs comprise double stranded RNA that can specifically block expression of a target gene. "RNA interference" or "RNAi" is a term initially applied to a phenomenon observed in plants and worms where double-stranded RNA (dsRNA) blocks gene expression in a specific and post-transcriptional manner. Without being bound by theory, RNAi appears to involve mRNA degradation, however the biochemical mechanisms are currently an active area of research. Despite some mystery regarding the mechanism of action, RNAi provides a useful method of inhibiting gene expression in vitro or in vivo.

As used herein, the term "dsRNA" refers to siRNA molecules, or other RNA molecules including a double stranded feature and able to be processed to siRNA in cells, such as hairpin RNA moieties.

The term "loss-of-function," as it refers to genes inhibited by the subject RNAi method, refers to a diminishment in the level of expression of a gene when compared to the level in the absence of RNAi constructs.

As used herein, the phrase "mediates RNAi" refers to (indicates) the ability to distinguish which RNAs are to be degraded by the RNAi process, e.g.,

degradation occurs in a sequence-specific manner rather than by a sequence-independent dsRNA response, e.g., a PKR response.

As used herein, the term "RNAi construct" is a generic term used throughout the specification to include small interfering RNAs (siRNAs), hairpin RNAs, and other RNA species which can be cleaved *in vivo* to form siRNAs. RNAi constructs herein also include expression vectors (also referred to as RNAi expression vectors) capable of giving rise to transcripts which form dsRNAs or hairpin RNAs in cells, and/or transcripts which can produce siRNAs *in vivo*.

"RNAi expression vector" (also referred to herein as a "dsRNA-encoding plasmid") refers to replicable nucleic acid constructs used to express (transcribe) RNA which produces siRNA moieties in the cell in which the construct is expressed. Such vectors include a transcriptional unit comprising an assembly of (1) genetic element(s) having a regulatory role in gene expression, for example, promoters, operators, or enhancers, operatively linked to (2) a "coding" sequence which is transcribed to produce a double-stranded RNA (two RNA moieties that anneal in the cell to form an siRNA, or a single hairpin RNA which can be processed to an siRNA), and (3) appropriate transcription initiation and termination sequences. The choice of promoter and other regulatory elements generally varies according to the intended host cell. In general, expression vectors of utility in recombinant DNA techniques are often in the form of "plasmids" which refer to circular double stranded DNA loops which, in their vector form are not bound to the chromosome. In the present specification, "plasmid" and "vector" are used interchangeably as the plasmid is the most commonly used form of vector. However, the invention is intended to include such other forms of expression vectors which serve equivalent functions and which become known in the art subsequently hereto.

The RNAi constructs contain a nucleotide sequence that hybridizes under physiologic conditions of the cell to the nucleotide sequence of at least a portion of the mRNA transcript for the gene to be inhibited (i.e., the "target" gene). The double-stranded RNA need only be sufficiently similar to natural RNA that it has the ability to mediate RNAi. Thus, the invention has the advantage of being able to tolerate sequence variations that might be expected due to genetic mutation, strain polymorphism or evolutionary divergence. The number of tolerated nucleotide

mismatches between the target sequence and the RNAi construct sequence is no more than 1 in 5 basepairs, or 1 in 10 basepairs, or 1 in 20 basepairs, or 1 in 50 basepairs. Mismatches in the center of the siRNA duplex are most critical and may essentially abolish cleavage of the target RNA. In contrast, nucleotides at the 3' end of the siRNA strand that is complementary to the target RNA do not significantly contribute to specificity of the target recognition.

Sequence identity may be optimized by sequence comparison and alignment algorithms known in the art (see Gribskov and Devereux, Sequence Analysis Primer, Stockton Press, 1991, and references cited therein) and calculating the percent difference between the nucleotide sequences by, for example, the Smith-Waterman algorithm as implemented in the BESTFIT software program using default parameters (e.g., University of Wisconsin Genetic Computing Group). Greater than 90% sequence identity, or even 100% sequence identity, between the inhibitory RNA and the portion of the target gene is preferred. Alternatively, the duplex region of the RNA may be defined functionally as a nucleotide sequence that is capable of hybridizing with a portion of the target gene transcript (e.g., 400 mM NaCl, 40 mM PIPES pH 6.4, 1 mM EDTA, 50 °C or 70 °C hybridization for 12-16 hours; followed by washing).

Production of RNAi constructs can be carried out by chemical synthetic methods or by recombinant nucleic acid techniques. Endogenous RNA polymerase of the treated cell may mediate transcription *in vivo*, or cloned RNA polymerase can be used for transcription *in vitro*. The RNAi constructs may include modifications to either the phosphate-sugar backbone or the nucleoside, e.g., to reduce susceptibility to cellular nucleases, improve bioavailability, improve formulation characteristics, and/or change other pharmacokinetic properties. For example, the phosphodiester linkages of natural RNA may be modified to include at least one of an nitrogen or sulfur heteroatom. Modifications in RNA structure may be tailored to allow specific genetic inhibition while avoiding a general response to dsRNA. Likewise, bases may be modified to block the activity of adenosine deaminase. The RNAi construct may be produced enzymatically or by partial/total organic synthesis, any modified ribonucleotide can be introduced by *in vitro* enzymatic or organic synthesis.

Methods of chemically modifying RNA molecules can be adapted for modifying RNAi constructs (see, for example, Heidenreich et al. (1997) *Nucleic Acids Res*, 25:776-780; Wilson et al. (1994) *J Mol Recog* 7:89-98; Chen et al. (1995) *Nucleic Acids Res* 23:2661-2668; Hirschbein et al. (1997) *Antisense Nucleic Acid Drug Dev* 7:55-61). Merely to illustrate, the backbone of an RNAi construct can be modified with phosphorothioates, phosphoramidate, phosphodithioates, chimeric methylphosphonate-phosphodiesters, peptide nucleic acids, 5-propynyl-pyrimidine containing oligomers or sugar modifications (e.g., 2'-substituted ribonucleosides, a-configuration).

The double-stranded structure may be formed by a single self-complementary RNA strand or two complementary RNA strands. RNA duplex formation may be initiated either inside or outside the cell. The RNA may be introduced in an amount which allows delivery of at least one copy per cell. Higher doses (e.g., at least 5, 10, 100, 500 or 1000 copies per cell) of double-stranded material may yield more effective inhibition, while lower doses may also be useful for specific applications. Inhibition is sequence-specific in that nucleotide sequences corresponding to the duplex region of the RNA are targeted for genetic inhibition.

In certain embodiments, the subject RNAi constructs are "small interfering RNAs" or "siRNAs." These nucleic acids are around 19-30 nucleotides in length, and even more preferably 21-23 nucleotides in length, e.g., corresponding in length to the fragments generated by nuclease "dicing" of longer double-stranded RNAs. The siRNAs are understood to recruit nuclease complexes and guide the complexes to the target mRNA by pairing to the specific sequences. As a result, the target mRNA is degraded by the nucleases in the protein complex. In a particular embodiment, the 21-23 nucleotides siRNA molecules comprise a 3' hydroxyl group.

The siRNA molecules of the present invention can be obtained using a number of techniques known to those of skill in the art. For example, the siRNA can be chemically synthesized or recombinantly produced using methods known in the art. For example, short sense and antisense RNA oligomers can be synthesized and annealed to form double-stranded RNA structures with 2-nucleotide overhangs at each end (Caplen, et al. (2001) *Proc Natl Acad Sci USA*, 98:9742-9747; Elbashir, et al. (2001) *EMBO J*, 20:6877-88). These double-stranded siRNA structures can then

be directly introduced to cells, either by passive uptake or a delivery system of choice, such as described below.

In certain embodiments, the siRNA constructs can be generated by processing of longer double-stranded RNAs, for example, in the presence of the enzyme dicer. In one embodiment, the Drosophila *in vitro* system is used. In this embodiment, dsRNA is combined with a soluble extract derived from Drosophila embryo, thereby producing a combination. The combination is maintained under conditions in which the dsRNA is processed to RNA molecules of about 21 to about 23 nucleotides.

The siRNA molecules can be purified using a number of techniques known to those of skill in the art. For example, gel electrophoresis can be used to purify siRNAs. Alternatively, non-denaturing methods, such as non-denaturing column chromatography, can be used to purify the siRNA. In addition, chromatography (e.g., size exclusion chromatography), glycerol gradient centrifugation, affinity purification with antibody can be used to purify siRNAs.

In certain preferred embodiments, at least one strand of the siRNA molecules has a 3' overhang from about 1 to about 6 nucleotides in length, though may be from 2 to 4 nucleotides in length. More preferably, the 3' overhangs are 1-3 nucleotides in length. In certain embodiments, one strand having a 3' overhang and the other strand being blunt-ended or also having an overhang. The length of the overhangs may be the same or different for each strand. In order to further enhance the stability of the siRNA, the 3' overhangs can be stabilized against degradation. In one embodiment, the RNA is stabilized by including purine nucleotides, such as adenosine or guanosine nucleotides. Alternatively, substitution of pyrimidine nucleotides by modified analogues, e.g., substitution of uridine nucleotide 3' overhangs by 2'-deoxythyinidine is tolerated and does not affect the efficiency of RNAi. The absence of a 2' hydroxyl significantly enhances the nuclease resistance of the overhang in tissue culture medium and may be beneficial *in vivo*.

In other embodiments, the RNAi construct is in the form of a long double-stranded RNA. In certain embodiments, the RNAi construct is at least 25, 50, 100, 200, 300 or 400 bases. In certain embodiments, the RNAi construct is 400-800 bases in length. The double-stranded RNAs are digested intracellularly, e.g., to produce

siRNA sequences in the cell. However, use of long double-stranded RNAs *in vivo* is not always practical, presumably because of deleterious effects which may be caused by the sequence-independent dsRNA response. In such embodiments, the use of local delivery systems and/or agents which reduce the effects of interferon or PKR are preferred.

In certain embodiments, the RNAi construct is in the form of a hairpin structure (named as hairpin RNA). The hairpin RNAs can be synthesized exogenously or can be formed by transcribing from RNA polymerase III promoters *in vivo*. Examples of making and using such hairpin RNAs for gene silencing in mammalian cells are described in, for example, Paddison et al., *Genes Dev*, 2002, 16:948-58; McCaffrey et al., *Nature*, 2002, 418:38-9; McManus et al., *RNA*, 2002, 8:842-50; Yu et al., *Proc Natl Acad Sci U S A*, 2002, 99:6047-52). Preferably, such hairpin RNAs are engineered in cells or in an animal to ensure continuous and stable suppression of a desired gene. It is known in the art that siRNAs can be produced by processing a hairpin RNA in the cell.

In yet other embodiments, a plasmid is used to deliver the double-stranded RNA, e.g., as a transcriptional product. In such embodiments, the plasmid is designed to include a "coding sequence" for each of the sense and antisense strands of the RNAi construct. The coding sequences can be the same sequence, e.g., flanked by inverted promoters, or can be two separate sequences each under transcriptional control of separate promoters. After the coding sequence is transcribed, the complementary RNA transcripts base-pair to form the double-stranded RNA.

PCT application WO01/77350 describes an exemplary vector for bi-directional transcription of a transgene to yield both sense and antisense RNA transcripts of the same transgene in a eukaryotic cell. Accordingly, in certain embodiments, the present invention provides a recombinant vector having the following unique characteristics: it comprises a viral replicon having two overlapping transcription units arranged in an opposing orientation and flanking a transgene for an RNAi construct of interest, wherein the two overlapping transcription units yield both sense and antisense RNA transcripts from the same transgene fragment in a host cell.

RNAi constructs can comprise either long stretches of double stranded RNA identical or substantially identical to the target nucleic acid sequence or short stretches of double stranded RNA identical to substantially identical to only a region of the target nucleic acid sequence. Exemplary methods of making and delivering either long or short RNAi constructs can be found, for example, in WO01/68836 and WO01/75164.

Exemplary RNAi constructs that specifically recognize a particular gene, or a particular family of genes can be selected using methodology outlined in detail above with respect to the selection of antisense oligonucleotide. Similarly, methods of delivery RNAi constructs include the methods for delivery antisense oligonucleotides outlined in detail above.

Ribozymes molecules designed to catalytically cleave an mRNA transcripts can also be used to prevent translation of mRNA (See, e.g., PCT International Publication WO90/11364, published October 4, 1990; Sarver et al., 1990, Science 247:1222-1225 and U.S. Patent No. 5,093,246). While ribozymes that cleave mRNA at site-specific recognition sequences can be used to destroy particular mRNAs, the use of hammerhead ribozymes is preferred. Hammerhead ribozymes cleave mRNAs at locations dictated by flanking regions that form complementary base pairs with the target mRNA. The sole requirement is that the target mRNA have the following sequence of two bases: 5'-UG-3'. The construction and production of hammerhead ribozymes is well known in the art and is described more fully in Haseloff and Gerlach, 1988, Nature, 334:585-591.

The ribozymes of the present invention also include RNA endoribonucleases (hereinafter "Cech-type ribozymes") such as the one which occurs naturally in Tetrahymena thermophila (known as the IVS, or L-19 IVS RNA) and which has been extensively described by Thomas Cech and collaborators (Zaug, et al., 1984, Science, 224:574-578; Zaug and Cech, 1986, Science, 231:470-475; Zaug, et al., 1986, Nature, 324:429-433; published International patent application No. WO88/04300 by University Patents Inc.; Been and Cech, 1986, Cell, 47:207-216). The Cech-type ribozymes have an eight base pair active site that hybridizes to a target RNA sequence whereafter cleavage of the target RNA takes place. The

invention encompasses those Cech-type ribozymes that target eight base-pair active site sequences.

As in the antisense approach, the ribozymes can be composed of modified oligonucleotides (e.g., for improved stability, targeting, etc.) and can be delivered to cells in vitro or in vivo. A preferred method of delivery involves using a DNA construct "encoding" the ribozyme under the control of a strong constitutive pol III or pol II promoter, so that transfected cells will produce sufficient quantities of the ribozyme to destroy targeted messages and inhibit translation. Because ribozymes unlike antisense molecules, are catalytic, a lower intracellular concentration is required for efficiency.

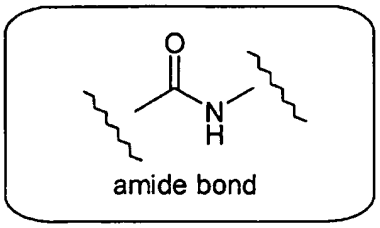
In other embodiments, the invention contemplates the use of peptidomimetics (herein referred to interchangeably as a peptide mimetics) to mimic the bioactivity of an embryonic stem cell specific protein. Such peptide mimetics can be identified by the screening methods of the invention and can be used to promote the embryonic stem cell phenotype in a cell.

Peptidomimetics are compounds based on, or derived from, peptides and proteins. The peptidomimetics of the present invention can be obtained by structural modification of the amino acid sequence of a known embryonic stem cell specific marker using unnatural amino acids, conformational restraints, isosteric replacement, and the like. The subject peptidomimetics constitute the continuum of structural space between peptides and non-peptide synthetic structures.

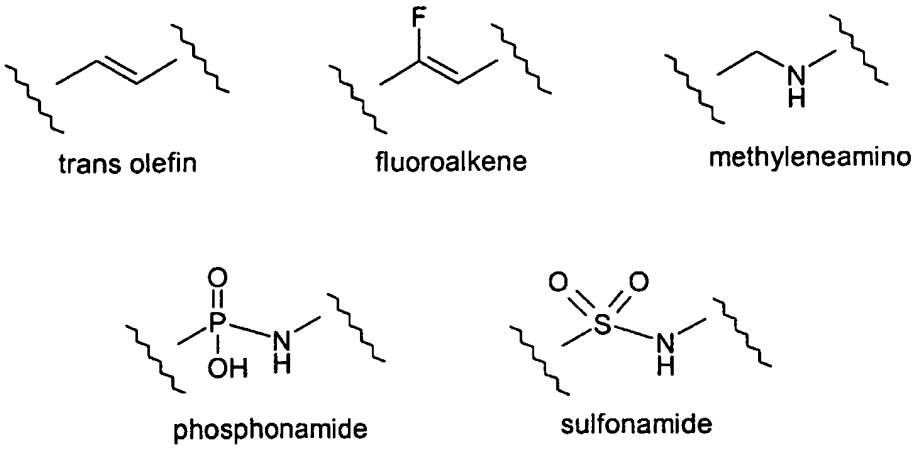
Exemplary peptidomimetics can have such attributes as being non-hydrolyzable (e.g., increased stability against proteases or other physiological conditions which degrade the corresponding peptide), having increased specificity and/or potency, and having increased cell permeability for intracellular localization. For illustrative purposes, peptide analogs of the present invention can be generated using, for example, benzodiazepines (e.g., see Freidinger et al. in Peptides: Chemistry and Biology, G.R. Marshall ed., ESCOM Publisher: Leiden, Netherlands, 1988), substituted gama lactam rings (Garvey et al. in *Peptides: Chemistry and Biology*, G.R. Marshall ed., ESCOM Publisher: Leiden, Netherlands, 1988, p123), C-7 mimics (Huffman et al. in *Peptides: Chemistry and Biologyy*, G.R. Marshall ed., ESCOM Publisher: Leiden, Netherlands, 1988, p. 105), keto-methylene

pseudopeptides (Ewenson et al. (1986) *J Med Chem* 29:295; and Ewenson et al. in *Peptides: Structure and Function* (Proceedings of the 9th American Peptide Symposium) Pierce Chemical Co. Rockland, IL, 1985), β-turn dipeptide cores (Nagai et al. (1985) *Tetrahedron Lett* 26:647; and Sato et al. (1986) *J Chem Soc Perkin Trans* 1:1231), β-aminoalcohols (Gordon et al. (1985) *Biochem Biophys Res Commun*126:419; and Dann et al. (1986) *Biochem Biophys Res Commun* 134:71), diaminoketones (Natarajan et al. (1984) *Biochem Biophys Res Commun* 124:141), and methyleneamino-modifed (Roark et al. in *Peptides: Chemistry and Biology*, G.R. Marshall ed., ESCOM Publisher: Leiden, Netherlands, 1988, p134). Also, see generally, Session III: Analytic and synthetic methods, in in *Peptides: Chemistry and Biology*, G.R. Marshall ed., ESCOM Publisher: Leiden, Netherlands, 1988)

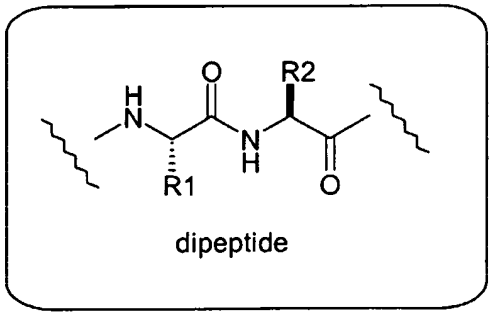
In addition to a variety of sidechain replacements which can be carried out to generate the subject peptidomimetics, the present invention specifically contemplates the use of conformationally restrained mimics of peptide secondary structure. Numerous surrogates have been developed for the amide bond of peptides. Frequently exploited surrogates for the amide bond include the following groups (i) trans-olefins, (ii) fluoroalkene, (iii) methyleneamino, (iv) phosphonamides, and (v) sulfonamides.



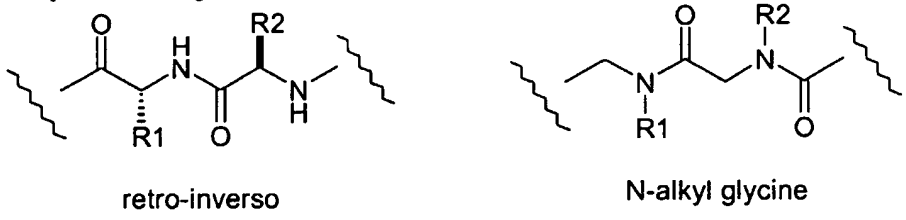
Examples of Surrogates



Additionally, peptidomimietics based on more substantial modifications of the backbone of a peptide can be used. Peptidomimetics which fall in this category include (i) retro-inverso analogs, and (ii) N-alkyl glycine analogs (so-called peptoids).

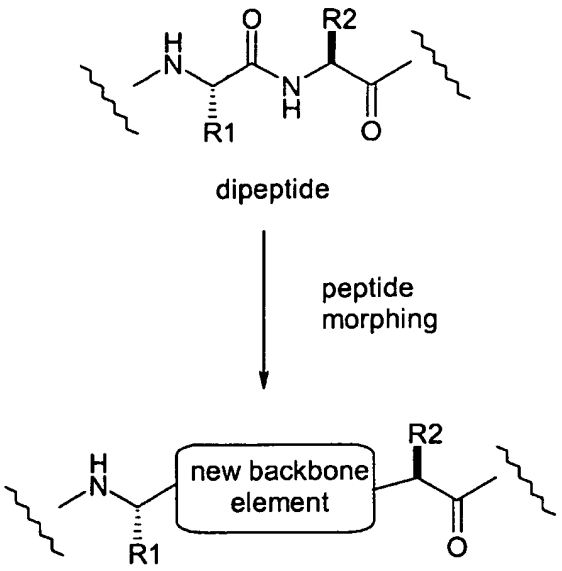


Examples of analogs



Furthermore, the methods of combinatorial chemistry are being brought to bear, e.g., PCT publication WO 99/48897, on the development of new peptidomimetics. For example, one embodiment of a so-called "peptide morphing"

strategy focuses on the random generation of a library of peptide analogs that comprise a wide range of peptide bond substitutes.



In an exemplary embodiment, the peptidomimetic can be derived as a retro-inverso analog of the peptide. Retro-inverso analogs can be made according to the methods known in the art, such as that described by the Sisto et al. U.S. Patent 4,522,752. As a general guide, sites which are most susceptible to proteolysis are typically altered, with less susceptible amide linkages being optional for mimetic switching The final product, or intermediates thereof, can be purified by HPLC.

In another illustrative embodiment, the peptidomimetic can be derived as a retro-enatio analog of the peptide. Retro-enantio analogs such as this can be synthesized using commercially available D-amino acids (or analogs thereof) and standard solid- or solution-phase peptide-synthesis techniques. For example, in a preferred solid-phase synthesis method, a suitably amino-protected (t-butyloxycarbonyl, Boc) residue (or analog thereof) is covalently bound to a solid support such as chloromethyl resin. The resin is washed with dichloromethane (DCM), and the BOC protecting group removed by treatment with TFA in DCM. The resin is washed and neutralized, and the next Boc-protected D-amino acid is introduced by coupling with diisopropylcarbodiimide. The resin is again washed, and the cycle repeated for each of the remaining amino acids in turn. When synthesis of the

protected retro-enantio peptide is complete, the protecting groups are removed and the peptide cleaved from the solid support by treatment with hydrofluoric acid/anisole/dimethyl sulfide/thioanisole. The final product is purified by HPLC to yield the pure retro-enantio analog.

In still another illustrative embodiment, trans-olefin derivatives can be made for any of the subject polypeptides. A trans olefin analog can be synthesized according to the method of Y.K. Shue et al. (1987) *Tetrahedron Letters* 28:3225 and also according to other methods known in the art. It will be appreciated that variations in the cited procedure, or other procedures available, may be necessary according to the nature of the reagent used.

It is further possible to couple the pseudodipeptides synthesized by the above method to other pseudodipeptides, to make peptide analogs with several olefinic functionalities in place of amide functionalities.

Still another classes of peptidomimetic derivatives include phosphonate derivatives. The synthesis of such phosphonate derivatives can be adapted from known synthesis schemes. See, for example, Loots et al. in *Peptides: Chemistry and Biology*, (Escom Science Publishers, Leiden, 1988, p. 118); Petrillo et al. in *Peptides: Structure and Function* (Proceedings of the 9th American Peptide Symposium, Pierce Chemical Co. Rockland, IL, 1985).

Many other peptidomimetic structures are known in the art and can be readily adapted for use in designing embryonic stem cell peptide mimetics. To illustrate, the peptidomimetic may incorporate the 1-azabicyclo[4.3.0]nonane surrogate (see Kim et al. (1997) J. Org. Chem. *62*:2847), or an *N*-acyl piperazic acid (see Xi et al. (1998) J. Am. Chem. Soc. *120*:80), or a 2-substituted piperazine moiety as a constrained amino acid analogue (see Williams et al. (1996) J. Med. Chem. *39*:1345-1348). In still other embodiments, certain amino acid residues can be replaced with aryl and bi-aryl moieties, e.g., monocyclic or bicyclic aromatic or heteroaromatic nucleus, or a biaromatic, aromatic, heteroaromatic, or biheteroaromatic nucleus.

The subject peptidomimetics can be optimized by, e.g., combinatorial synthesis techniques combined with high throughput screening techniques, and furthermore can be tested to ensure that the peptidomimetic retains one or more of

the biological activities of the native polypeptide. Note the invention contemplates peptide mimetics designed to all or a portion of the stem cell specific amino acid sequences represented in SEQ ID NO: 2, 4, 6, 8, or 10.

Exemplification

The invention now being generally described, it will be more readily understood by reference to the following examples which are included merely for purposes of illustration of certain aspects and embodiments of the present invention, and are not intended to limit the invention.

Example 1: Human Embryonic Stem Cells are Most Closely Related to Cells of the Reproductive System and Hematopoietic Progenitor Cells

We used gene profiling analysis to compare global gene expression across a large number of cell types. The purpose of this analysis was two fold. Firstly, by comparing the gene expression profile of a number of human embryonic stem cell lines, we confirmed that independently isolated embryonic stem cells are substantially similar not only in terms of their overt physical properties but also on a molecular level. Secondly, this analysis aimed to ascertain the degree of similarity with respect to global gene expression of many diverse cell types.

METHODS: Culture of embryonic stem cells and embryoid bodies - Human embryonic stem cells were cultured as previously described with minor modifications (Itskovitz-Eldor et al. (2000) *Molecular Medicine* **6**: 88-95; Schuldiner et al. (2000) *PNAS* **97**: 11307-11312). Briefly, undifferentiated human embryonic stem cells from either of two independently identified embryonic stem cells lines (X or Y) were cultured in 80% KnockOut™ DMEM medium (Gibco-BRL) on a mitomycin-C treated mouse embryonic fibroblast (MEF) feeder layer. The culture medium was supplemented with the following: 20% KnockOut™ SR (a serum free formulation manufactured by Gibco-BRL), 1 mM glutamine (Gibco-BRL), 0.1 mM β-mercaptoethanol (Sigma), 1% non-essential amino acids (Gibco-BRL), 50 units/ml penicillin, 50 μg/ml streptomycin, and 4 ng/ml basic fibroblast growth factor (bFGF). To reduce the presence of feeder cells in the culture, human embryonic stem cells were grown on 0.1% gelatin (Merck) coated plates for a one passage. High density cultures of undifferentiated embryonic stem cells (approx.

10^6 cells/plate) were trypsinized and used for either RNA extraction or for embryoid body formation. In instances where the cells were used for embryoid body formation, embryonic stem cells were allowed to aggregate in suspension on plastic Petri dishes. Embryoid bodies were collected and analyzed following 2, 10, 20, or 30 days of aggregation and culture.

DNA micro-array analysis – Total RNA was extracted from populations of embryonic stem cells, embryoid bodies, or differentiated cell types. RNA extraction was performed using the manufacturer's protocol for optimal RNA preparation for array analysis (Affymetrix). Global gene expression in each sample was analyzed using Affymetrix DNA chip micro-arrays U95 and U133. Briefly, U95 contains over 12,000 human expressed sequences, and was used to compare gene expression in human embryonic stem cells and embryoid bodies to gene expression in approximately 40 different tissue types. U133 is an improved version of U95 and contains nearly 45,000 probe sets representing 33,000 human genes. U133 was used to compare global gene expression of human ES cells (cell line X) and embryoid bodies cultures for 2, 10 and 30 days. Hybridization to the DNA micro-arrays, washing and scanning were all performed in accordance with the manufacturer's instructions. The following programs were used to analyze the results: CLUSTER, TREEVIEW, and GENE SRING. The following NCBI databases were consulted: Unigene, Entrez, Blast, LocusLink, and Aceview. Additionally, the SOURCE database from Stanford University and the BLAT database from the University of California at Santa Cruz were used.

Figure 1 summarizes the results of this global gene expression analysis. Figure 1A summarizes the gene expression analysis comparing global gene expression in ES cells, EBs, and other somatic cell types. The results provided in figure 1A are displayed as a branched tree, wherein the degree of similarity is represented by branch height. We note that the 3 ES cell samples cluster together, and cluster relatively close to EBs. Of the other cell types examined, ES cells and EBs appear to be most closely related to reproductive and hematopoietic cell types.

To further explore the relationship between various ES cell lines and EBs at various point during differentiation, we assessed similarities in global gene expression across these cell types. The results are summarized in Figure 1B, and

confirm that there is very little variation in gene expression across different ES cell lines, and that different ES cell lines are more closely related to each other than to EBs.

Example 2: Identification of Novel ES Cell Markers

Despite the tremendous interest in the field of embryonic stem cells, to date our understanding of the molecular nature of these cells is limited. The study of the molecular nature of embryonic stem cells has been hampered by the absence of embryonic stem cell markers. Preferably, ES cell markers are those that are expressed relatively specifically in ES cells. Alternatively, when such markers are also expressed in other tissues, the other tissues should be easily distinguishable from ES cells such that the ability to properly identify an ES cell is not compromised.

To identify ES cell specific genes, we performed a pair wise comparison of gene expression between ES cells and 30 day old embryoid bodies using the U133 Affymetrix micro-array. The average signal value was calculated for each probe, and ordered by the ES/EB ratio. Probes for which this ratio was greater than 20 were analyzed in further detail. Using this initial criteria for selecting probes for further consideration, we focused on 73 sequences.

Further examination of the 73 candidate sequences was conducted. Further analysis included database searches to eliminate sequences whose pattern of expression was already known.

Figure 2 provides RT-PCR results which confirm that five sequences identified based on micro-array analysis are in fact differentially expressed between ES cells and EBs. These five sequences (OCT4 and 4 novel sequences) represent markers of ES cells based on their robust and relatively specific expression in ES cells in comparison to EBs. We note that these markers are also expressed in germ cells.

Briefly, the identification of OCT4 (referred to in Figure 2 as Hs.2860 and in the sequence listing as SEQ ID NO: 1), a known marker of ES cells, confirmed that the outlined methodology could successfully be employed to identify ES cell markers. However, the function of the other four sequences remains unknown. Sequence analysis predicts that all four genes encode transcription factors, however

the significance of this observation is not clear. The four novel embryonic stem cell markers are as follows: Hs.67624 (which is referred to in the sequence listing as SEQ ID NO: 3), Hs.143925 (which is referred to in the sequence listing as SEQ ID NO: 5 and is not shown in Figure 2), Hs.86154 (which is referred to in the sequence listing as SEQ ID NO: 7), and Hs. 189095 (which is referred to in the sequence listing as SEQ ID NO: 9).

METHODS – RT-PCR – Total RNA was extracted from either ES cells or from day 20 embryoid bodies using standard methods. 1 μg of RNA was reverse transcribed by random hexamer priming using the EZ-First Strand cDNA Synthesis Kit (Biological Industries). Amplification was performed on the cDNA using Takara Ex Taq™ in the presence of 1X Ex Taq™ buffer, 200 μM each dNTPs, and 2.5 mM Mg^{2+}. The PCR conditions included a first step of 3 minutes at 94 °C, a second step of 20-30 cycles of 30 seconds at 94 °C, a 30 second annealing step at 62-64 °C, a 45 second 72 °C step, and a final step of 5 minutes at 72 °C. We note that GPDH is used as a control.

The following specific primer pairs were used to amplify the ES specific genes. For each combination of primers, we have also provided the temperature at which the 30 second annealing step was performed and the size of the fragment amplified by the particular primer pair.

OCT4:Annealing temp = 64 °C; fragment size = 637 bp.

5' primer – gatcctcggacctggctaag (SEQ ID NO: 11)

3' primer – ctctcactcggttctcgatac (SEQ ID NO: 12)

SEQ ID NO: 3: Annealing temp = 64 °C, fragment size = 649 bp.

5' primer – ggttctctgactgactccttc (SEQ ID NO: 13)
3' primer – gctcctggcagctctttattc (SEQ ID NO: 14)

SEQ ID NO: 5: Annealing temp = 62 °C, fragment size = 491 bp.

5' primer – ggtgccatgactcggatcg (SEQ ID NO: 15)
3' primer – ctcacagtacttgctgtagg (SEQ ID NO: 16)

SEQ ID NO: 7: Annealing temp = 64 °C, fragment size = 523 bp.

5' primer – caccagaataagctgcacatg (SEQ ID NO: 17)

3' primer – cctgagatacatggcagtgc (SEQ ID NO: 18)

SEQ ID NO: 9: Annealing temp = 62 °C, fragment size = 464 bp.

5' primer – caggaatttgtggcggagag (SEQ ID NO: 19)
3' primer – cctgtgacagtccttactgc (SEQ ID NO: 20)

GPDH: Annealing temp = 62 °C, fragment size = 302 bp.

5' primer – agccacatcgctcagacacc (SEQ ID NO: 21)
3' primer – gtactcagcggccagcatcg (SEQ ID NO: 22)

Final products were assessed by gel electrophoresis on a 2% ethidium bromide (EtBr) stained agarose gel, and the identity of the amplified products were verified by sequencing.

Example 3: Analysis of ES Specific Markers

Figures 3-6 provide analysis of the ES specific markers identified.

Figure 3 provides analysis of the ES specific nucleic acid represented in SEQ ID NO: 3. The largest open reading frame identified for this ES specific marker is provided in SEQ ID NO: 4.

Figure 4 provides analysis of the ES specific nucleic acid represented in SEQ ID NO: 5. The amino acid sequence for a 3' open reading frame is provided in SEQ ID NO: 6.

Figure 5 provides analysis of the ES specific nucleic acid represented in SEQ ID NO: 7. The mRNA encodes a protein similar to the C. elegans lin-28 RNA binding protein, and the protein encoded by SEQ ID NO: 7 is represented in SEQ ID NO: 8. Analysis of this protein revealed the presence of domains consistent with a role in RNA binding and binding to single stranded DNA including zinc knuckle domain. This is the first observation indicating a role for a lin-28 related protein in ES cell fate.

Figure 6 provides analysis of the ES specific nucleic acid represented in SEQ ID NO: 9. The mRNA encodes a protein known as Sall4, and the protein encoded by SEQ ID NO: 9 is represented in SEQ ID NO: 10. Sall4 belongs to the family of C2H2 zinc finger domain containing proteins that are thought to function as transcription factors. Consistent with the domain structure observed for other

members of this family, analysis of SEQ ID NO: 10 revealed the presence of seven C2H2 zinc finger domains.

Example 4: Patterns of Gene Expression in ES Cells and Embryoid Bodies Over Time Recapitulate Patterns of Gene Expression During Development

To begin to assess whether differentiation of embryonic stem cells via embryoid bodies of various ages recapitulates differentiation of tissues during development, we began to examine the temporal expression of particular genes. The temporal expression of these genes was compared to that which is observed during development. Briefly, we have observed that the temporal pattern of gene expression for several genes involved in nodal signaling recapitulates that observed during development.

Figure 7 summarizes these results as observed by micro-array profiling (Figure 7A) and by RT-PCR (Figure 7B). As during normal development, nodal was expressed in ES cells and in day 2 EBs, however, the expression of nodal decreased in day 10 EBs and is virtually undetectable in day 30 EBs. LeftyA and LeftyB were bearly detectable (if at all) in ES cells. Their expression increased dramatically in day two EBs, decreased in day 10 EBs, and was virtually undetectable in day 30 EBs. Pitx2 is undetectable in ES cells, day 2 EBs, and day 10 EBs. Expression of Pitx2 increased in later EBs and was robustly observed in day 30 EBs.

METHODS – RT-PCR – Total RNA was extracted from either ES cells, day 2 EBs, day 10 EBs, or day 30 EBs using standard methods. 1 μg of RNA was reverse transcribed by random hexamer priming using the EZ-First Strand cDNA Synthesis Kit (Biological Industries). Amplification was performed on the cDNA using Takara Ex Taq™ in the presence of 1X Ex Taq™ buffer, 200 μM each dNTPs, and 2.5 mM Mg^{2+}. The PCR conditions included a first step of 3 minutes at 94 °C, a second step of 20-30 cycles of 30 seconds at 94 °C, a 30 second annealing step at 62-64 °C, a 45 second 72 °C step, and a final step of 5 minutes at 72 °C. We note that GPDH (also known as GAPDH) was used as a control.

The following specific primer pairs were used to amplify each genes. For each combination of primers, we have also provided the temperature at which the

second annealing step was performed and the size of the fragment amplified by the particular primer pair.

GPDH: Annealing temp = 62 °C, fragment size = 302 bp.

5' primer – agccacatcgctcagacacc (SEQ ID NO: 21)
3' primer – gtactcagcggccagcatcg (SEQ ID NO: 22)

Nodal: Annealing temp = 64 °C, fragment size = 535 bp.

5' primer – ggcagaagatgtggcagtgg (SEQ ID NO: 23)
3' primer – caagtgatgtcgacggtgc (SEQ ID NO: 24)

LeftyA: Annealing temp = 62 °C, fragment size = 435 bp.

5' primer – ctggacctcagggactatg (SEQ ID NO: 25)
3' primer – gaccacctcttatgcacacg (SEQ ID NO: 26)

LeftyB: Annealing temp = 62 °C, fragment size = 520 bp.

5' primer – cacaagctggtccgctttg (SEQ ID NO: 27)
3' primer – caggtaccctcgaacacttc (SEQ ID NO: 28)

Pitx2: Annealing temp = 62 °C, fragment size = 307 bp.

5' primer – gtggaccaaccttacggaag (SEQ ID NO: 29)
3' primer – catgctcatggacgagatag (SEQ ID NO: 30)

Final products were assessed by gel electrophoresis on a 2% ethidium bromide (EtBr) stained agarose gel, and the identity of the amplified products were verified by sequencing.

Example 5: Gene Profiling Analysis of Embryonic Stem Cells and Embryoid Bodies

As outlined in Examples 1 and 2, the four embryonic stem cell specific genes described herein were identified via a two step process. However, the initial step also provided a great deal of information relevant to the present invention. Table 1 summarizes the results of our initial analysis. These markers were preferentially expressed in embryonic stem cells in comparison to embryoid bodies. Accordingly, examining one or more of the markers provided in Table 1 is useful in identifying embryonic stem cells, and distinguishing undifferentiated embryonic stem cells from differentiated cells types. Examples of differentiated cell types are the diverse populations of differentiated cell types found in embryoid bodies.

In light of the wealth of expression data collected from gene profiling analysis of embryonic stem cells and embryoid bodies (summarized in Table 1), the present invention provides methods of identifying and characterizing an embryonic stem cell by examining the expression of one or more of the markers provided in Table 1. For example, the invention contemplates examining at least 1, 2, 3, 4, 5, 10,15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 66, or 67 of the markers identified in Table 1. A cell that preferentially expresses one or more of these markers, in comparison to expression of the markers in embryoid bodies, is a candidate embryonic stem cell.

SEQUENCES DESCRIBED IN THE APPLICATION:

Human Oct4 nucleic acid sequence (SEQ ID NO: 1)

gtagtcctttgttacatgcatgagtcagtgaacagggaatgggtgaatgacatttgtgggtaggttatttctagaagtta
ggtgggcagctcggaaggcagatgcacttctacagactattccttggggccacacgtaggttcttgaatcccgaatggaa
aggggagattgataactggtgtgtttatgttcttacaagtcttctgccttttaaaatccagtcccaggacatcaaagctc
tgcagaaagaactcgagcaatttgccaagctcctgaagcagaagaggatcaccctgggatatacacaggccgatgtgg
ggctcaccctgggggttctatttgggaaggtattcagccaaacgaccatctgccgctttgaggctctgcagcttagcttcaa
gaacatgtgtaagctgcggcccttgctgcagaagtgggtggaggaagctgacaacaatgaaaatcttcaggagatatgc
a
aagcagaaaccctcgtgcaggcccgaaagagaaagcgaaccagtatcgagaaccgagtgagaggcaacctggaga
atttgttcctgcagtgcccgaaacccacactgcagcagatcagccacatcgcccagcagcttgggctcgagaaggatgt
ggtccgagtgtggttctgtaaccggcgtagtcctttgttacatgcatgagtcagtgaacagggaatgggtgaatgacatttg
tggg
taggttatttctagaagttaggtgggcagctcggaaggcagatgcacttctacagactattccttggggccacacgtagg
ttcttgaatcccgaatggaaaggggagattgataactggtgtgtttatgttcttacaagtcttctgccttttaaaatcca
gtcccaggacatcaaagctctgcagaaagaactcgagcaatttgccaagctcctgaagcagaagaggatcaccctggg
at
atacacaggccgatgtggggctcaccctgggggttctatttgggaaggtattcagccaaacgaccatctgccgctttgag
gctctgcagcttagcttcaagaacatgtgtaagctgcggcccttgctgcagaagtgggtggaggaagctgacaacaatg
a
aaatcttcaggagatatgcaaagcagaaaccctcgtgcaggcccgaaagagaaagcgaaccagtatcgagaaccgag
tgagaggcaacctggagaatttgttcctgcagtgcccgaaacccacactgcagcagatcagccacatcgcccagcagct
tgggctcgagaaggatgtggtccgagtgtggttctgtaaccggcgccagaagggcaagcgatcaagcagcgactatgc
acaacgagaggattttgaggctgctgggtctcctttctcaggggaccagtgtcctttcctctggccccagggccccatttt
ggtg
ccccaggctatgggagccctcacttcactgcactgtactcctcggtccctttccctgaggggaagcctttccccctgtc
tctgtcaccactctgggctctcccttgcattcaaactgaggtgcctgcctgcccttctaggaatgggggacaggggagg
ggaggagctagggaaagaaaacctggagtttgtgccagggtttttggattaagttcttcattcactaaggaaggaattgg
gaacacaaagggtgggggcaggggagtttggggcaactggttggagggaaggtgaagttcaatgatgctcttgatttta
a
tcccacatcatgtatcacttttttcttaaataaagaagcttgggacacagtagataga (SEQ ID NO: 1)

Human Oct4 amino acid sequence (SEQ ID NO: 2)

MHFYRLFLGATRRFLNPEWKGEIDNWCVYVLTSLLPFKIQSQDIKALQKELE
QFAKLLKQKRITLGYTQADVGLTLGVLFGKVFSQTTICRFEALQLSFKNMCK
LRPLLQKWVEEADNNENLQEICKAETLVQARKRKRTSIENRVRGNLENLFL
QCPKPTLQQISHIAQQLGLEKDVVRVWFCNRRQKGKRSSSDYAQREDFEAA
GSPFSGGPVSFPLAPGPHFGAPGYGSPHFTALYSSVPFPEGEAFPPVSVTTLGS
PLHSN (SEQ ID NO: 2)

Nucleic acid sequence for the ES marker described in Figure 3 (SEQ ID NO: 3)

CTGGCTCAAAAAGCACCCCCACTGAGCACCTTGCGACCCCCCGCTCCTAC
CCGCCAGAGAACAAACCCCCTTTGACTGTAATTTTCCTTTACCTAACCAA
ATCCTATAAAACGGCCCCACCCTTATCTCCCTTCGCTGACTCTCTTTTCGG
ACTCAGCCCGCCTGCACCCAGGTGAAATAAACAGCCTCGTTGCTCACAC
AAAGCCTGTTTGGTGGTCTCTTCACACGGACGCGCATGAAATTTGGTGCC
GTGACTCGGATCGGGGGACCTCCCTTGGGAGATCAATCCCCTGTCCTCCT
GCTCTTTGCTCCGTGAGAAAGATCCACCTACGACCTCAGGTCCTCAGACC
AACCAGCCCAAGAAACATCTCACCAATTTCAAATCCGGTAAGCGGCCTC
TTTTTACTCTGTTCTCCAACCTCCCTCACTATCCCTCAACCTCTTTCTCCTT
TCAATCTTGGCGCCACACTTCAATCTCTCCCTTCTCTTAATTTCAATTCCT
TTCATTCTCTGGTAGAGACAAAAGAGACATGTTTTATCCGTGAACCCAAA
ACTCCGGCGCCGGTCACGGACTGGGAAGGCAGTCTTCCCTTGGTGTTTAA
TCATTGCAGGGACGCCTCTCTGATTTCACGTTTCAGACCACGCAGGGATG
CCTGCCTTGGTCCTTCACCCTTAGCGGCAAGTCCCGCTTTCCTGGGGCAG
GGGCAAGTACCCCTCAACCCCTTCTCCTTCACCCTTAGCGGCAAGTCCCG
CTTTTCTGGGGCAGGGGCAAGTACCCCTCAACCCCTTCTCCTTCACCCTT
AGCAGCAAGTCCCGCTTTCCTAGGGGGCAAGAACCCCCCAATCGCTTAT
TTTCACGCCCCAACAGAAACCCCCACCCCTTCTCCGTGTCTCTACTCTTTT
CTCTGGGCTTGCCTCCTTCACTATGGGCAAGCTTCCACCTTCCATTCCTTT
CTTCTCCCTTAGCATGTATTCTTAAGAACTTAAAATCTCTTCAATTCTCAC
CTGACCTAAAATCTAAGCGTCTTATTTTCTTCTGCAATGCCACTTGACCC
CAATACAAACTCAACAGTAGTTCCAAATAGCCAGAAAATGGCACTTTCA
ATTTTTCCACCCTACAAGATCTAAATAATTCTTGGCGTAAAATGGGCAAA
TGGTGTGAGGTGCCTGACGTCCAGGCATTCTTTTACACATCAGTCCCTTC
CTAGTCTCTGTGCCCAGTGCAACTCGTCCCAAATCTTCCTTCTTTCCCTCC
CGCCTGTCCCCTCAGTACCAACCCCAAGCGTCACTGAGTCTTTCTAATCT
TCCTTTTCTACAGACCCATCTGACCTCTCCCTTCCTCCCCAGGCTGCTCCT
TGCCAGGCCGAGCTAGGTCCCAATTCTTCCTCAGCCTCACACAAGAACTT
CCAAACGCCTGAACTGTAGCAGCCAGACGTTTCTCCAGAACCTCCTCCCC
CAGGAACTTGCTACACATGCCGGAAATCTGGCCACTGGGCCAAGGAACG
CCCGCAGCCCGGGATTCCTCCTAAGCCGCGTCCCATCTGTGTGGGACCCC
ACTGAAAATCGGACTGTTCAACTCACCTGGCAGCCACTCCCAGAGCTCCT
GGAACTCTGGCCCAAGGTTCTCTGACTGACTCCTTCTTGGCTTACTGGCT
GAAGACTGACGCTGCCTGATCGCCTCAGAAGCCCCGCAGACCATCATGG
ACGCCGAGCTTTAGCCCGCCTGCACCCAGGTGAAATAAACAGCCTTGTT
GCTCACACAAAGCCTGTTTGGTGGTCTCTTCACACAGACGCGCATGAAA
GGGAAGACATACAAAAACAAGGTATCTGAGGTAGGTACTACTGAGACA
GCCAGGTGGGAAGGACTCCTTGGCAAAACTCCAACCAGCCTGTACACTG

GGAGGAATGTGCACTGGGATGGAGCCATAGAAGTTTGTGTCGTTTGCAG
TGGGGAGGAGCCTGGTCCCTCCTCTTCCTGTGAGGAACCTGGAATTCAAT
CTGTGAGGTTGTTCTGGAGATGTTCTGGGGAGACTGCATTAAACACAGCT
TCGCACCATTGAATAAACTCAGCAACAAGCCAATGCATAAAAGTAATCT
ATGCTTCAGGTCACAGAAGCTTCAAGGGGAAAAAAACAGAATACTCTAG
GGCCATTGTTCACAAACTCATCTGAAAACATCCTGGAAAAATTTTCCCAA
ACACATGGAAAGAAAGAGAGGAAAAAAGAAGATATCTGAATAATGTGG
ACTAGAATAAAGAGCTGCCAGGAGCTGTTTATTTAAAAACAGTACTTTCT
TCTCTGGCTGAGTCCCTGGTATTCTCTGCTGCAATCTGTAGCTGTAGAAT
TTTGAAGAATGCAATTAAATTCAAATGGTTTGATGAGTAATAT (SEQ ID NO: 3)

DNA Sequence of BF223023 mRNA with Repetitive sequences masked (N) with Repeatmasker <http://repeatmasker.genome.washington.edu>:

NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNAANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNN

NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNCGAGCTNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNAAGGGAAGACATACAAAAACAAGGTATCTGAGGTA
GGTACTACNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTGTTCTGGAGATGTTCT
GGGGAGACTGCATTAAACACAGCTTCGCACCATTGAATAAACTCAGCAA
CAAGCCAATGCATAAAAGTAATCTATGCTTCAGGTCACAGAAGCTTCAA
GGGGAAAAAAACAGAATACTCTAGGGCCATTGTTCACAAACTCATCTGA
AAACATCCTGGAAAAATTTTCCCAAACACATGGAAAGAAAGAGAGGAA
AAAAGAAGATATCTGAATAATGTGGACTAGAATAAAGAGCTGCCAGGA
GCTGTTTATTTAAAAACAGTACTTTCTTCTCTGGCTGAGTCCCTGGTATTC
TCTGCTGCAATCTGTAGCTGTAGAATTTTGAAGAATGCAATTAAATTCAA
ATGGTTTGATGAGTAATAT

Sequence of largest ORF: (SEQ ID NO: 4)

PNPIKRPHPYLPSLTLFSDSARLHPGEINSLVAHTKPVWWSLHTDAHEIWCR
DSDRGTSLGRSIPCPPALCSVRKIHLRPQVLRPTSPRNISPISNPVSGLFLLCSP
TSLTIPQPLSPFNLGATLQSLPSLNFNSFHSLVETKETCFIREPKTPAPVTDWE
GSLPLVFNHCRDASLISRFRPRRDACLGPSPLAASPAFLGQGQVPLNPFSFTL
SGKSRFSGAGASTPQPLLLHP (SEQ ID NO: 4)

Nucleic acid sequence for the ES marker described in Figure 4 (SEQ ID NO: 5)

GGCTGACTCTCTTTTCGGACTCAGCCCGCCTGCACCCAGGTGAAATAAAC
AGCCTTGTTGCTCACACAAAGCCTGTTTGGTGGTCTCTTCACACAAACGC
GCATGAAATTTGGTGCCATGACTCGGATCGGGGTACCTCCCTTGGGAGA
TCAATCCCCAGTCCTCCTGCTCTTTGCTCCGTGAGAAAGATCTACCTAGG
ACCTCAGGTCCTCAGACTGACCAGCCCAAGGAACATCTCACCAATTTCA
AATCTGGAACGCGCATGAAAAAACCAAACAAACAAAAAAATTCTTTTGG
TAGCAGAATAAAAAAACAAAAAAAAGGACTTTTTCTTCTGGACTGAACT
ATATTTAAATCTCAAAGGATGGACATCTCACAACCTTCCTACAGCAAGTA
CTGTGAGAGCTGCATCTTGTCCCACTGGATGGTCTTCAGAGACAATAATA
CATAATGGAGCTGTCATCTCCTATGATAACAATGCCTTCTTCTGGATACC
TCCTGAAGGACCTGCCTGAGGCTCTTTCACTCCATGAAAAGGTGCGCTGC
TTTCCCTTTGCCTTCCGCCATTATTGAAAGTTTCCTGAGGCCTCCCCAGCC
ATGCTACCTGTACGACCTGTGAAACCATAAGCCAAAAAAGATACTAGCG
CCAGTCTGGCAGGGGCCTTTTCTAGTCTCAAGCACGCTGAGCAGTCCTAC
ACCTTGCCTTCTGAGAGAGAAGAGGACAGTCCTCAGCCTCCTGAAGCCT
GGCAAGGGTGCCTTAGTGAGGACATCATCTGCAACTCCCAGATGGATAG
ACGAGGCCATGGAGAGAGGAGAACACCAACCCATGAGTGACAAGTGCC
TGTCATTGTCAATGATGCTGCAAGCACCACGATGCCCCTTCTTAGCAGGG
ACCCAGTGGGCCTTACCAGCTTCATTATTTTCTAGGCATGATCGCATCCT
GTCTCCCATTTTTGCTAAAATTCTATCTAGAGAAAGAGCTCATTTTTAGC
TAAAAAAAAAAAAATCCTGTTTTATCTGCTAAAAAATCACATAAAATAA

CTATTGGACTGTCAAAAAAAAAAAAAAAAAAAAAAAAAAAA (SEQ ID NO: 5)

DNA Sequence of BC026300 mRNA with Repetitive sequences masked (N) with Repeatmasker <http://repeatmasker.genome.washington.edu>:

GNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNAANNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNCATGAAAAAACCAAACAAACA
AAAAAATTCTTTTGGTAGCAGAATAAAAAAACAAAAAAAAGGACTTTTT
CTTCTGGACTGAACTATATTTAAATCTCAAAGGATGGACATCTCACAACC
TTCCTACAGCAAGTACTGTGAGAGCTGNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNAAAAGATACTAGCGCCAGTCTGGCAGGGGCCTT
TTCTAGTCTCAAGCACGCTGAGCAGTCCTACACCTTGCCTTCTGAGAGAG
AAGAGGACAGTCCTCAGCCTCCTGAAGCCTGGCAAGGGTGCCTTAGTGA
GGACATCATCTGCAACTCCCAGATGGATAGACGAGGCCATGGAGAGAGG
AGAACACCAACCCATGAGTGACAAGTGCCTGTCATTGTCAATGATGCTG
CAAGCACCACGATGCCCCTTCTTAGCAGGGACCCAGTGGGCCTTACCAG
CTTCATTATTTTCTAGGCATGATCGCATCCTGTCTCCCATTTTTGCTAAAA
TTCTATCTAGAGAAAGAGCTCATTTTTAGCTAAAAAAAAAAAAAATCCTG
TTTTATCTGCTAAAAAATCACATAAAATAACTATTGGACTGTCNNNNNN
NNNNNNNNNNNNNNNNNNNNNNN

Amino Acid Sequence of 3' exon ORF (SEQ ID NO: 6)
MERGEHQPMSDKCLSLSMMLQAPRCPFLAGTQWALPASLFSRHDRILSPIFA
KILSRERAHF

Nucleic acid sequence for the ES marker described in Figure 5 (SEQ ID NO: 7)

GGGGCCACGGGCTCAGCCGACGACCATGGGCTCCGTGTCCAACCAGCAG
TTTGCAGGTGGCTGCGCCAAGGCGGCAGAAGAGGCGCCCGAGGAGGCG
CCGGAGGACGCGGCCCGGGCGGCGGACGAGCCTCAGCTGCTGCACGGTG
CGGGCATCTGTAAGTGGTTCAACGTGCGCATGGGGTTCGGCTTCCTGTCC
ATGACCGCCCGCGCCGGGGTCGCGCTCGACCCCCCAGTGGATGTCTTTGT
GCACCAGAGTAAGCTGCACATGGAAGGGTTCCGGAGCTTGAAGGAGGGT
GAGGCAGTGGAGTTCACCTTTAAGAAGTCAGCCAAGGGTCTGGAATCCA
TCCGTGTCACCGGACCTGGTGGAGTATTCTGTATTGGGAGTGAGAGGCG
GCCAAAAGGAAAGAGCATGCAGAAGCGCAGATCAAAAGGAGACAGGTG
CTACAACTGTGGAGGTCTAGATCATCATGCCAAGGAATGCAAGCTGCCA
CCCCAGCCCAAGAAGTGCCACTTCTGCCAGAGCATCAGCCATATGGTAG

CCTCATGTCCGCTGAAGGCCCAGCAGGGCCCTAGTGCACAGGGAAAGCC
AACCTACTTTCGAGAGGAAGAAGAAGAAATCCACAGCCCTACCCTGCTC
CCGGAGGCACAGAATTGAGCCACAATGGGTGGGGGCTATTCTTTTGCTA
TCAGGAAGTTTTGAGGAGCAGGCAGAGTGGAGAAAGTGGGAATAGGGT
GCATTGGGGCTAGTTGGCACTGCCATGTATCTCAGGCTTGGGTTCACACC
ATCACCCTTTCTTCCCTCTAGGTGGGGGGAAAGGGTGAGTCAAAGGAAC
TCCAACCATGCTCTGTCCAAATGCAAGTGAGGGTTCTGGGGGCAACCAG
GAGGGGGGAATCACCCTACAACCTGCATATTTTGAGTCTCCATCCCCAG
AATTTCCAGCTTTTGAAAGTGGCCTGGATAGGGAAGTTGTTTTCCTTTTA
AAGAAGGATATATAATAATTCCCATGCCAGAGTGAAATGATTAAGTATA
AGACCAGATTCATGGAGCCAAGCCACTACATTCTGTGGAAGGAGATCTC
TCAGGAGTAAGCATTGTTTTTTTTTCACATCTTGTATCCTCATACCCACTT
TTGGGATAGGGTGCTGGCAGCTGTCCCAAGCAATGGGTAATGATGATGG
CAAAAAGGGTGTTTGGGGGAACAGCTGCAGACCTGCTGCTCTATGCTCA
CCCCCGCCCCATTCTGGGCCAATGTGATTTTATTTATTTGCTCCCTTGGAT
ACTGCACCTTGGGTCCCACTTTCTCCAGGATGCCAACTGCACTAGCTGTG
TGCGAATGACGTATCTTGTGCATTTTAACTTTTTTTCCTTAATATAAATAT
TCTGGTTTTGTATTTTTGTATATTTTAATCTAAGGCCCTCATTTCCTGCAC
TGTGTTCTCAGGTACATGAGCAATCTCAGGGATAGCCAGCAGCAGCTCC
AGGTCTGCGCAGCAGGAATTACTTTTTGTTGTTTTTGCCACCGTGGAGAG
CAACTATTTGGAGTGCACAGCCTATTGAACTACCTCATTTTTGCCAATAA
GAGCTGGCTTTTCTGCCATAGTGTCCTCTTGAAACCCCCTCTGCCTTGAA
AATGTTTTATGGGAGACTAGGTTTTAACTGGGTGGCCCCATGACTTGATT
GCCTTCTACTGGAAGATTGGGAATTAGTCTAAACAGGAAATGGTGGTAC
ACAGAGGCTAGGAGAGGCTGGGCCCGGTGAAAAGGCCAGAGAGCAAGC
CAAGATTAGGTGAGGGTTGTCTAATCCTATGGCACAGGACGTGCTTTAC
ATCTCCAGATCTGTTCTTCACCAGATTAGGTTAGGCCTACCATGTGCCAC
AGGGTGTGTGTGTGTTTGTAAAACTAGAGTTGCTAAGGATAAGTTTAAA
GACCAATACCCCTGTACTTAATCCTGTGCTGTCGAGGGATGGATATATGA
AGTAAGGTGAGATCCTTAACCTTTCAAAATTTTCGGGTTCCAGGGAGAC
ACACAAGCGAGGGTTTTGTGGTGCCTGGAGCCTGTGTCCTGCCCTGCTAC
AGTAGTGATTAATAGTGTCATGGTAGCTAAAGGAGAAAAAGGGGGTTTC
GTTTACACGCTGTGAGATCACCGCAAACCTACCTTACTGTGTTGAAACGG
GACAAATGCAATAGAACGCATTGGGTGGTGTGTGTCTGATCCTGGGTTCT
TGTCTCCCCTAAATGCTGCCCCCCAAGTTACTGTATTTGTCTGGGCTTTGT
AGGACTTCACTACGTTGATTGCTAGGTGGCCTAGTTTGTGTAAATATAAT
GTATTGGTCTTTCTCCGTGTTCTTTGGGGGTTTTGTTTACAAACTTCTTTTT
GTATTGAGAGAAAAATAGCCAAAGCATCTTTGACAGAAGGTTCTGCACC
AGGCAAAAAGATCTGAAACATTAGTTTGGGGGGCCCTCTTCTTAAAGGG
GGGATCTTGAACCATCCTTTCTTTTGTATTCCCCTTCCCCTATTACCTATT
AGACCAGATCTTCTGTCCTAAAAACTTGTCTTCTACCCTGCCCTCTTTTCT
GTTCACCCCCAAAAGAAAACTTACACACCCACACACATACACATTTCAT
GCTTGGAGTGTCTCCACAACTCTTAAATGATGTATGCAAAAATACTGAA
GCTAGGAAAACCCTCCGTCCCTTGTTCCCAACCTCCTAAGTCAAGACCAT
TACCATTTCTTTCTTTCTTTTTTTTTTTTTTTTTAAAGTGGAGTCTCGCTGT
GTCACCCAGGCAGAGGTTGCAGTGAGCTGAGATCGCACCACTGCACTCC
AGCCTGGTTACAGAGCGAGACTCTGTCTCAAACAAAACAAAACAAAACA
AAAACACACTACTGTATTTTGGATGGATCAAACCTCCTTAATTTTAATTT

CTAATCCTAAAGTAAAGAGATGCAATTGGGGGCCTTCCATGTAGAAAGT
GGGGTCAGGAGGCCAAGAAAGGGAATATGAATGTATATCCAAGTCACTC
AGGAACTTTTATGCAGGTGCTAGAAACTTTATGTCAAAGTGGCCACAAG
ATTGTTTAATAGGAGACGAACGAATGTAACTCCATGTTTACTGCTAAAA
ACCAAAGCTTTGTGTAAAATCTTGAATTTATGGGGCGGGAGGGTAGGAA
AGCCTGTACCTGTCTGTTTTTTTCCTGATCCTTTTCCCTCATTCCTGAACT
GCAGGAGACTGAGCCCCTTTGGGCTTTGGTGACCCCATCACTGGGGTGT
GTTTATTTGATGGTTGATTTTGCTGTACTGGGTACTTCCTTTCCCATTTTC
TAATCATTTTTTTAACACAAGCTGACTCTTCCCTTCCCTTCTCCTTTCCCTG
GGAAAATACAATGAATAAATAAAGACTTATTGGTACGC (SEQ ID NO: 7)

Protein Sequence: (SEQ ID NO: 8)

MGSVSNQQFAGGCAKAAEEAPEEAPEDAARAADEPQLLHGAGICKWFNVR
MGFGFLSMTARAGVALDPPVDVFVHQSKLHMEGFRSLKEGEAVEFTFKKS
AKGLESIRVTGPGGVFCIGSERRPKGKSMQKRRSKGDRCYNCGGLDHHAKE
CKLPPQPKKCHFCQSISHMVASCPLKAQQGPSAQGKPTYFREEEEEIHSPTLL
PEAQN (SEQ ID NO: 8)

Nucleic acid sequence for the ES marker described in Figure 6 (SEQ ID NO: 9)

AACTCCAGGAATTTGTGGCGGAGAGGGCAAATAACTGCGGCTCTCCCGG
CGCCCCGATGCTCGCACCATGTCGAGGCGCAAGCAGGCGAAACCCCAGC
ACATCAACTCGGAGGAGGACCAGGGCGAGCAGCAGCCGCAGCAGCAGA
CCCCGGAGTTTGCAGATGCGGCCCCAGCGGCGCCCGCGGCGGGGGAGCT
GGGTGCTCCAGTGAACCACCCAGGGAATGACGAGGTGGCGAGTGAGGA
TGAAGCCACAGTAAAGCGGCTTCGTCGGGAGGAGACGCACGTCTGTGAG
AAATGCTGTGCGGAGTTCTTCAGCATCTCTGAGTTCCTGGAACATAAGAA
AAATTGCACTAAAAATCCACCTGTCCTCATCATGAATGACAGCGAGGGG
CCTGTGCCTTCAGAAGACTTCTCCGGAGCTGTACTGAGCCACCAGCCCAC
CAGTCCCGGCAGTAAGGACTGTCACAGGGAGAATGGCGGCAGCTCAGA
GGACATGAAGGAGAAGCCGGATGCGGAGTCTGTGGTGTACCTAAAGAC
AGAGACAGCCCTGCCACCCACCCCCCAGGACATAAGCTATTTAGCCAAA
GGCAAAGTGGCCAACACTAATGTGACCTTGCAGGCACTACGGGGCACCA
AGGTGGCGGTGAATCAGCGGAGCGCGGATGCACTCCCTGCCCCCGTGCC
TGGTGCCAACAGCATCCCGTGGGTCCTCGAGCAGATCTTGTGTCTGCAGC
AGCAGCAGCTACAGCAGATCCAGCTCACCGAGCAGATCCGCATCCAGGT
GAACATGTGGGCCTCCCACGCCCTCCACTCAAGCGGGGCAGGGGCCGAC
ACTCTGAAGACCTTGGGCAGCCACATGTCTCAGCAGGTTTCTGCAGCTGT
GGCTTTGCTCAGCCAGAAAGCTGGAAGCCAAGGTCTGTCTCTGGATGCC
TTGAAACAAGCCAAGCTACCTCACGCCAACATCCCTTCTGCCACCAGCTC
CCTGTCCCCAGGGCTGGCACCCTTCACTCTGAAGCCGGATGGGACCCGG
GTGCTCCCGAACGTCATGTCCCGCCTCCCGAGCGCTTTGCTTCCTCAGGC
CCCGGGCTCGGTGCTCTTCCAGAGCCCTTTCTCCACTGTGGCGCTAGACA
CATCCAAGAAAGGGAAGGGGAAGCCACCGAACATCTCCGCGGTGGATG
TCAAACCCAAAGACGAGGCGGCCCTCTACAAGCACAAGTGTAAGTACTG
TAGCAAGGTTTTTGGGACTGATAGCTCCTTGCAGATCCACCTCCGCTCCC
ACACTGGAGAGAGACCCTTCGTGTGCTCTGTCTGTGGTCATCGCTTCACC

ACCAAGGGCAACCTCAAGGTGCACTTTCACCGACATCCCCAGGTGAAGG
CAAACCCCCAGCTGTTTGCCGAGTTCCAGGACAAAGTGGCGGCCGGCAA
TGGCATCCCCTATGCACTCTCTGTACCTGACCCCATAGATGAACCGAGTC
TTTCTTTAGACAGCAAACCTGTCCTTGTAACCACCTCTGTAGGGCTACCT
CAGAATCTTTCTTCGGGGACTAATCCCAAGGACCTCACGGGTGGCTCCTT
GCCCGGTGACCTGCAGCCTGGGCCTTCTCCAGAAAGTGAGGGTGGACCC
ACACTCCCTGGGGTGGGACCAAACTATAATTCCCCAAGGGCTGGTGGCT
TCCAAGGGAGTGGGACCCCTGAGCCAGGGTCAGAGACCCTGAAATTGCA
GCAGTTGGTGGAGAACATTGACAAGGCCACCACTGATCCCAACGAATGT
CTCATTTGCCACCGAGTCTTAAGCTGTCAGAGCTCCCTCAAGATGCATTA
TCGCACCCACACCGGGGAGAGACCGTTCCAGTGTAAGATCTGTGGCCGA
GCCTTTTCTACCAAAGGTAACCTGAAGACACACCTTGGGGTTCACCGAA
CCAACACATCCATTAAGACGCAGCATTCGTGCCCCATCTGCCAGAAGAA
GTTCACTAATGCCGTGATGCTGCAGCAACATATTCGGATGCACATGGGC
GGTCAGATTCCCAACACGCCCCTGCCAGAGAATCCCTGTGACTTTACGG
GTTCTGAGCCAATGACCGTGGGTGAGAACGGCAGCACCGGCGCTATCTG
CCATGATGATGTCATCGAAAGCATCGATGTAGAGGAAGTCAGCTCCCAG
GAGGCTCCCAGCAGCTCCTCCAAGGTCCCCACGCCTCTTCCCAGCATCCA
CTCGGCATCACCCACGCTAGGGTTTGCCATGATGGCTTCCTTAGATGCCC
CAGGGAAAGTGGGTCCTGCCCCTTTTAACCTGCAGCGCCAGGGCAGCAG
AGAAAACGGTTCCGTGGAGAGCGATGGCTTGACCAACGACTCATCCTCG
CTGATGGGAGACCAGGAGTATCAGAGCCGAAGCCCAGATATCCTGGAAA
CCACATCCTTCCAGGCACTCTCCCCGGCCAATAGTCAAGCCGAAAGCAT
CAAGTCAAAGTCTCCCGATGCTGGGAGCAAAGCAGAGAGCTCCGAGAAC
AGCCGCACTGAGATGGAAGGTCGGAGCAGTCTCCCTTCCACGTTTATCC
GAGCCCCGCCGACCTATGTCAAGGTTGAAGTTCCTGGCACATTTGTGGG
ACCCTCGACATTGTCCCCAGGGATGACCCCTTTGTTAGCAGCCCAGCCAC
GCCGACAGGCCAAGCAACATGGCTGCACACGGTGTGGGAAGAACTTCTC
GTCTGCTAGCGCTCTTCAGATCCACGAGCGGACTCACACTGGAGAGAAG
CCTTTTGTGTGCAACATTTGTGGGCGAGCTTTTACCACCAAAGGCAACTT
AAAGGTTCACTACATGACACACGGGGCGAACAATAACTCAGCCCGCCGT
GGAAGGAAGTTGGCCATCGAGAACACCATGGCTCTGTTAGGTACGGACG
GAAAAAGAGTCTCAGAAATCTTTCCCAAGGAAATCCTGGCCCCTTCAGT
GAATGTGGACCCTGTTGTGTGGAACCAGTACACCAGCATGCTCAATGGC
GGTCTGGCCGTGAAGACCAATGAGATCTCTGTGATCCAGAGTGGGGGGG
TTCCTACCCTCCCGGTTTCCTTGGGGGCCACCTCCGTTGTGAATAACGCC
ACTGTCTCCAAGATGGATGGCTCCCAGTCGGGTATCAGTGCAGATGTGG
AAAAACCAAGTGCTACTGACGGCGTTCCCAAACACCAGTTTCCTCACTTC
CTGGAAGAAAACAAGATTGCGGTCAGCTAAGGGAGAACTTGCGTGGAA
GGAGCAATGCAGACACAGTGAAATCTCTAGAATCTGCTTTGTTTTGTAAG
AACTCATCTCCTCCTGTTTTCTTTTTCTTACTGATATGCAAATGATGTTTA
CTACGTTGGTTGTGACCACAACCTCAGGCAAGTGCTACAATCACGATTGT
TGCTATGCTGCTTTGCAAAAAGTTGAAAAAATAAAAAAAAAAATGCATAC
CAAAAC (SEQ ID NO: 9)

Protein Sequence (SEQ ID NO: 10)

MSRRKQAKPQHINSEEDQGEQQPQQQTPEFADAAPAAPAAGELGAPVNHP
GNDEVASEDEATVKRLRREETHVCEKCCAEFFSISEFLEHKKNCTKNPPVLI

MNDSEGPVPSEDFSGAVLSHQPTSPGSKDCHRENGGSSEDMKEKPDAESVV
YLKTETALPPTPQDISYLAKGKVANTNVTLQALRGTKVAVNQRSADALPAP
VPGANSIPWVLEQILCLQQQQLQQIQLTEQIRIQVNMWASHALHSSGAGADT
LKTLGSHMSQQVSAAVALLSQKAGSQGLSLDALKQAKLPHANIPSATSSLSP
GLAPFTLKPDGTRVLPNVMSRLPSALLPQAPGSVLFQSPFSTVALDTSKKGK
GKPPNISAVDVKPKDEAALYKHKCKYCSKVFGTDSSLQIHLRSHTGERPFVC
SVCGHRFTTKGNLKVHFHRHPQVKANPQLFAEFQDKVAAGNGIPYALSVPD
PIDEPSLSLDSKPVLVTTSVGLPQNLSSGTNPKDLTGGSLPGDLQPGPSPESEG
GPTLPGVGPNYNSPRAGGFQGSGTPEPGSETLKLQQLVENIDKATTDPNECLI
CHRVLSCQSSLKMHYRTHTGERPFQCKICGRAFSTKGNLKTHLGVHRTNTSI
KTQHSCPICQKKFTNAVMLQQHIRMHMGGQIPNTPLPENPCDFTGSEPMTV
GENGSTGAICHDDVIESIDVEEVSSQEAPSSSSKVPTPLPSIHSASPTLGFAMM
ASLDAPGKVGPAPFNLQRQGSRENGSVESDGLTNDSSSLMGDQEYQSRSPDI
LETTSFQALSPANSQAESIKSKSPDAGSKAESSENSRTEMEGRSSLPSTFIRAP
PTYVKVEVPGTFVGPSTLSPGMTPLLAAQPRRQAKQHGCTRCGKNFSSASA
LQIHERTHTGEKPFVCNICGRAFTTKGNLKVHYMTHGANNNSARRGRKLAI
ENTMALLGTDGKRVSEIFPKEILAPSVNVDPVVWNQYTSMLNGGLAVKTNE
ISVIQSGGVPTLPVSLGATSVVNNATVSKMDGSQSGISADVEKPSATDGVPK
HQFPHFLEENKIAVSES (SEQ ID NO: 10)

Additional References

Goldstein et al. (2002) *Developmental Dynamics* **225**: 80-86.
Schuldiner et al. (2001) *Brain Research* **913**: 201-205.
Pan et al. (2002) *Cell Research* **12**: 321-329.
Eiges and Benvenisty (2002) *FEBS Letters* **529**: 135-141.
Schuldiner et al. (2000) *PNAS* **97**: 11307-11312.
Shamblott et al. (2001) *PNAS* **98**: 113-118.
Thomson et al. (1998) *Science* **282**: 1145-1147.
Pesce and Scholer (2001) *Stem Cells* **19**: 271-278.
Pesce and Scholer (2000) *Mol Rep Dev* **55**: 452-457.
Nichols et al. (1998) *Cell* **95**: 379-391.
Niwa et al. (2000) *Nature Genetics* **24**: 372-376.
Hansis et al. (2000) *Mol Hum Reprod* **6**: 999-1004.
Palumbo et al. (2002) *Journal of Pathology* **196** : 467-477.
Monk. (2001) *Oncogene* **20**: 8085-8091.
Kohlhase et al. (2002) *Human Molecular Genetics* **11**: 2979-2987.
Al-Baradie et al. (2002) *American Journal of Human Genetics* **71**: 1195-1199.
Reubinoff et al. (2000) *Nat Biotechnol* **18**: 399-404.
Schuldiner and Benvenisty. (2001) *Recent Res. Devel. Mol. Cell. Biol* **2**: 223-231.
Itskovitz-Eldor et al. (2000) *Mol Med* **6**: 88-95.
Reubinoff et al. (2001) *Nat Biotechnol* **19**: 1134-40.
Carpenter et al. (2001) *Exp Neurol* **172**: 383-97.
Zhang et al. (2001) *Nat Biotechnol* **19**: 1129-33.
Kehat et al. (2001) *J Clin Invest* **108**: 407-14.
Mummery et al. (2002) *J Anat* **200**: 233-42.
Kaufman et al. (2001) *Proc Natl Acad Sci U S A* **98**: 10716-21.
Eiges et al. (2001) *Curr Biol* **11**: 514-8.

Henderson et al. (2002) *Stem Cells* **20:** 329-37.
Ramalho-Santos et al. (2002) *Science* **298:** 597-600.
Ivanova et al. (2002) *Science* **298:** 601-4.
Eisen et al. (1998) *Proc Natl Acad Sci U S A* **95:** 14863-8.
Hamada et al. (2002) *Nat Rev Genet* **3:** 103-13.
Leahy et al. (1999) *J Exp Zool* **284:** 67-81.
Pelton et al. (2002) *J Cell Sci* **115:** 329-39.
Gillespie and Uversky. (2000) *Biochim Biophys Acta* **1480:** 41-56.
Crouch. (1998) *Biochim Biophys Acta* **1408:** 278-89.
WO00/70021
WO02/10347
WO02/061033

All publications, patents and patent applications are herein incorporated by reference in their entirety to the same extent as if each individual publication, patent or patent application was specifically and individually indicated to be incorporated by reference in its entirety.

Equivalents

Those skilled in the art will recognize, or be able to ascertain using no more than routine experimentation, many equivalents to the specific embodiments of the invention described herein. Such equivalents are intended to be encompassed by the following claims.